FIXED ACTIVATION BOUNDS ARE NOT ENOUGH: A DISTORTION-MATCHED TRANSFORMER STUDY

Anonymous authors
Paper under double-blind review

ABSTRACT

There are increasing societal concerns regarding safety around AI. Safe AI requires accurate predictions that withstand adversarial perturbations. Fixed bounds on internal activations constrain representation magnitudes, but whether they improve decision robustness remains uncertain. Here we show that fixed, training-calibrated hard and smooth caps preserve clean accuracy without establishing the prespecified practical robustness benefit in binary DistilRoBERTa classification on the Internet Movie Database benchmark. Across 55 matched training runs, we compare five transformer locations while matching mean absolute intervention on training-only activations; every nontrivial hard-cap distortion admits a unique positive smooth scale. All ten capped configurations satisfy corrected clean balanced-accuracy equivalence within one percentage point. At the primary $0.05r$ embedding-space budget, none meets robustness-benefit criteria. Among 27 endpoint contrasts for caps with recorded activity, 15 establish corrected equivalence and 12 remain unresolved; the historically inactive smooth-logit comparator adds three equivalence decisions. At the less saturated $0.025r$ budget, a post-hoc paired analysis likewise finds no corrected positive cap-control difference: hard query/key capping is 5.56 percentage points worse, seven contrasts remain unresolved, and two are equivalent. A contextual FreeLB-style baseline confirms that the evaluator distinguishes a more robust trained system. Our findings show a clear boundary between fixed activation control and robustness gains.

1 INTRODUCTION

Bounding an internal representation gives direct control over its magnitude. Whether it improves the model's decisions is a separate question. Learned activation caps have been investigated for adversarial robustness (Sooksatra & Rivas, 2024), and rectified activations (ReAct) can improve out-of-distribution detection (Sun et al., 2021). Yet a bounded output range alone need not confer adversarial robustness (Kanai et al., 2021). We ask: *under a common calibration and training policy, do fixed activation bounds improve transformer classifier robustness, and does their location or operator make a practically important difference?*

A useful comparison must distinguish operator shape from intervention strength. Hard clipping leaves small coordinates unchanged, whereas a smooth cap contracts every nonzero coordinate in real arithmetic. We therefore match their mean absolute change on the same training-only sample rather than equate their numerical scales. Theorem 1 guarantees a unique smooth scale for every nontrivial target distortion, making the comparison well posed without promising robustness.

We evaluate an uncapped control and ten fixed capped configurations (55 matched runs) at five DistilRoBERTa sites on the binary Internet Movie Database (IMDb) benchmark. Bounds are calibrated on training activations and frozen before fine-tuning; learned or dynamic bounds are outside scope. Hard caps are attacked with backward pass differentiable approximation (BPDA), rather than credited for obstructing gradients (Athalye et al., 2018), and equivalence is tested against a specified margin rather than inferred from nonsignificance (Lakens, 2017). All ten caps preserve clean balanced accuracy (BA) within the margin, but none establishes the primary practical robustness benefit. Existing query-key normalization (QKNorm) and three-seed FreeLB-style comparisons provide context; only the latter occupies a substantially more robust regime (Henry et al., 2020; Zhu et al., 2020).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The contributions are threefold: a *formal distortion-matching construction*; a *controlled 55-run, five-location study* with matched seeds and adaptive evaluation; and *endpoint-specific evidence separating clean-utility preservation from robustness benefit across primary and diagnostic budgets*. Figure 1 explains the comparison, while Table 1 retains every configuration. The study thus distinguishes a controllable representation property from the robustness outcome that motivates it.

2 Related work

Magnitude control serves different objectives. Dynamic-Max-Value ReLU extends the rectified linear unit (ReLU) with learned, regularized nonnegative maxima in an added dense layer of image classifiers (Sooksatra & Rivas, 2024). Our fixed symmetric caps occupy native transformer locations instead. Other objectives differ: Parameterized Clipping Activation for Quantized Neural Networks (PACT) learns clipping for quantization (Choi et al., 2018), Dynamic Tanh replaces normalization (Zhu et al., 2025), and ReAct and activation shaping target out-of-distribution detection (Sun et al., 2021; Djurisic et al., 2023). Most directly, Kanai et al. (2021) distinguish ordinary tanh-bounded logits from a construction that also controls pre-logit norms. These precedents motivate testing what a bound achieves, rather than treating boundedness itself as novelty.

Location and training changes are not interchangeable. Query-key normalization (QKNorm) combines normalized projections with a learned attention scale (Henry et al., 2020). Robust optimization and FreeLB instead change the training objective (Madry et al., 2018; Zhu et al., 2020). Neither is equivalent to inserting a fixed coordinate cap. Outlier dimensions can also carry structure associated with token frequency and attention (Puccetti et al., 2022); large values are not necessarily expendable. Our distortion-matched design controls operator strength on the calibration sample while leaving gradients and subsequent engagement free to differ. Appendix D relates this design to additional attention, certification, and representation-attack work.

Evaluation must distinguish confidence from correctness. Saturation calls for adaptive attacks and stronger-search diagnostics (Athalye et al., 2018; Carlini et al., 2019; Croce & Hein, 2020). Miscalibration can interfere with natural language processing (NLP) attack search (Raina et al., 2024), while discrete TextFooler success and textual validity require separate assessment (Jin et al., 2020; Zhou et al., 2024). Practical equivalence likewise requires its own margin and tests (Lakens, 2017). These distinctions motivate our joint evaluation. We first define the intervention and the comparison supported by its mathematical properties.

3 Distortion-matched activation capping

Our construction fixes a hard threshold from an uncapped control, then matches the smooth operator's mean absolute intervention on the same calibration values. This separates operator shape from one defined measure of intervention strength.

3.1 Operators and native insertion points

For a scalar $x \in \mathbb{R}$, define

$$C_\tau(x) = \operatorname{sign}(x)\min\{|x|, \tau\}, \qquad S_s(x) = s\tanh(x/s), \qquad \tau, s > 0, \tag{1}$$

with $\operatorname{sign}(0) = 0$. Both maps act coordinatewise on tensors. Each configuration selects one operator at one of five locations; it does not cap all five locations simultaneously. A block-local choice is applied in all six transformer layers.

Embedding capping follows layer normalization (LN) of the embeddings and precedes embedding dropout. *Query/key capping* follows projection and head reshaping, before the scaled dot product; queries (Q) and keys (K) share a layer-specific bound. *Attention-output capping* follows head concatenation and the output projection, before branch dropout, residual addition, and LayerNorm. *Multilayer perceptron (MLP) capping* follows the Gaussian error linear unit (GELU) and precedes the feed-forward network (FFN) output projection. Finally, *centered-logit capping* acts on $\operatorname{ctr}(z) = z - K^{-1}\mathbf{1}\mathbf{1}^\top z$ before loss and softmax, where $z \in \mathbb{R}^K$, $\mathbf{1}$ is the all-ones vector, and

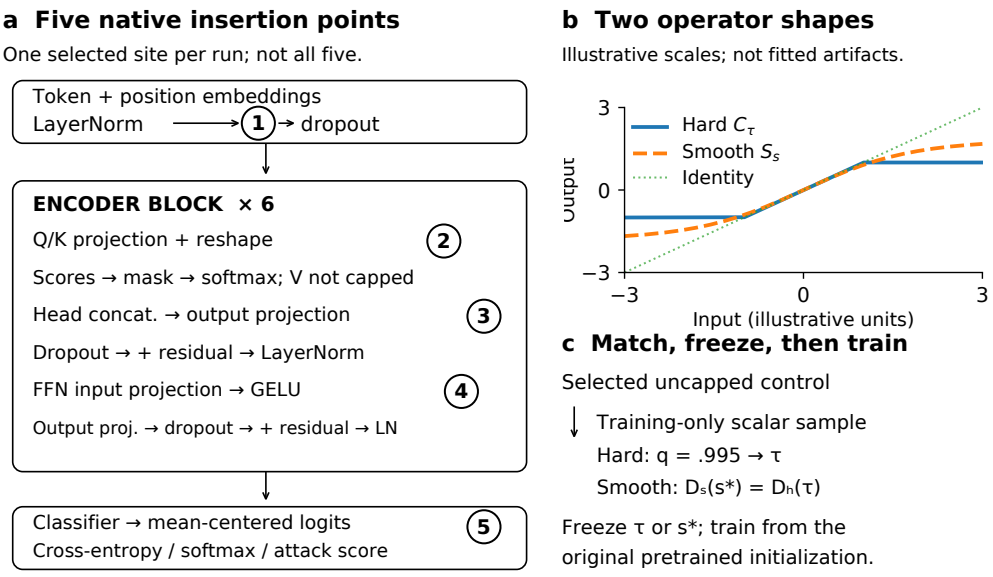

Figure 1: Where and how capping acts. (a) One of five numbered native sites is selected per configuration; block-local sites act in all six layers. Q/K is capped before masking, attention output before dropout and residual addition, and the MLP after GELU but before its output projection. (b) Curves illustrate operator shapes only; the displayed scales are not fitted experimental artifacts. (c) The selected uncapped control supplies training-only calibration values. Hard and smooth parameters are matched on those values, frozen, and used to train separate capped models from the original pretrained initialization.

$K = 2$ here. All configurations, including the control, use this centering; without a cap it leaves softmax probabilities and cross-entropy unchanged in real arithmetic. Appendix Table 2 maps these operations to the exact implementation identifiers and clarifies the residual and mask ordering. In particular, masks are applied after finite Q/K scores are formed; the cap never acts on a masking sentinel. Figure 1 places these interventions on the forward path and separates the illustrative operator shapes from the training-only calibration procedure. The fixed experimental comparison matches operator distortion, not the numerical heights drawn in the schematic.

3.2 Calibration and the matching theorem

For each model seed, we first train and select the uncapped control. In evaluation mode, it supplies a deterministic scalar sample $\mathcal{A} = (a_1, \ldots, a_N)$ from a balanced, training-only set of 2,048 reviews, excluding padding. Each seed and location has its own sample; block-local sites additionally distinguish layers, with query and key samples pooled within a layer. The hard threshold is the empirical $q = 0.995$ `higher` quantile of $|\mathcal{A}|$. The quantile is fixed, not selected for favorable robustness. Appendix B records the sampling and artifact conventions.

Let $(v)_+ = \max(v, 0)$ and define the two mean absolute distortions:

$$D_H(\tau) = \frac{1}{N}\sum_{i=1}^{N}(|a_i| - \tau)_+, \qquad D_S(s) = \frac{1}{N}\sum_{i=1}^{N}|a_i - s\tanh(a_i/s)|. \tag{2}$$

We seek $s^\star > 0$ with $D_S(s^\star) = D_H(\tau)$. The following result ensures that this is a well-defined comparison rather than an ambiguous scale search.

Theorem 1 (Unique distortion-matched scale)**.** *Let $a_1, \ldots, a_N \in \mathbb{R}$ with $N \geq 1$, and write $\mu = N^{-1}\sum_i |a_i|$. Assume $\mu > 0$. Then D_S is continuous and strictly decreasing on $(0, \infty)$, with $\lim_{s\downarrow 0} D_S(s) = \mu$ and $\lim_{s\to\infty} D_S(s) = 0$. Consequently, for every target $d \in (0, \mu)$ there is exactly one $s^\star > 0$ satisfying $D_S(s^\star) = d$.*

Proof. For $u > 0$, $0 < \tanh u < u$. Thus, by oddness, a sample term with magnitude $a = |a_i|$ equals $f_a(s) = a - s\tanh(a/s)$; f_0 is identically zero. For $a > 0$, put $u = a/s$. Then

$$f_a'(s) = -g(u), \qquad g(u) = \tanh u - u\,\mathrm{sech}^2 u. \tag{3}$$

Because $g(0) = 0$ and $g'(u) = 2u\,\mathrm{sech}^2 u\tanh u > 0$ for $u > 0$, every nonzero term decreases strictly. Their finite average is therefore continuous and strictly decreasing. As $s \downarrow 0$, $0 \le s\tanh(a/s) \le s$, so $f_a(s) \to a$. As $s \to \infty$, $s\tanh(a/s) = a\,\tanh(u)/u \to a$, so $f_a(s) \to 0$. Averaging gives the two limits. For any $d \in (0, \mu)$, those limits bracket d at finite positive scales; continuity supplies a solution and strict decrease makes it unique. □

For $\tau > 0$, positive hard distortion automatically lies below μ. Thus the theorem applies to every accepted nonzero-distortion artifact. Moreover, $\tau\tanh(a/\tau) < \min(a, \tau)$ for $a > 0$, so $D_S(\tau) > D_H(\tau)$ and the matched scale satisfies $s^\star > \tau$. Thus setting $s = \tau$ would overstate the smooth intervention. For small absolute targets, the large-scale expansion in Appendix A.1 shows that $s^\star$ can grow as the inverse square root of the target distortion; executed activity must still be measured rather than inferred from scale alone. Bisection uses the fixed relative-mismatch tolerance 10^{-6}; a zero-distortion artifact is invalid rather than assigned an arbitrary finite scale. Appendix A supplies the endpoint cases and numerical stopping rule.

The cap parameters are then fixed, and every capped model is trained anew from the original pretrained initialization, not the fitted control. Matching holds within the same seed, site, layer when applicable, and calibration sample. It does not equalize gradients, hit rates, tail shape, distortion across locations, or distortion after capped training. Those differences remain part of the study.

3.3 What the bounds do (and do not) guarantee

Write T_b for either operator at a fixed positive parameter b. For vectors $x, y \in \mathbb{R}^m$ and $1 \le p \le \infty$,

$$\|T_b(x)\|_\infty \le b, \qquad \|T_b(x) - T_b(y)\|_p \le \|x - y\|_p. \tag{4}$$

For Q/K capping, these local bounds also limit finite attention scores and their softmax concentration. Appendix A.2 proves the resulting probability and entropy bounds, and the corresponding centered-logit confidence ceiling. These are checkable local consequences: downstream transformations, residual paths, and class margins are not controlled by (4) alone.

For a fixed binary classifier, output-only capping makes this distinction particularly clear. Its centered logits are $(d/2, -d/2)$ with $d = z_1 - z_2$; a positive-scale hard or smooth cap preserves the sign of their difference and therefore the exact decision boundary, including ties (Proposition 3). Confidence can change without correcting a single decision. Training with the cap can nevertheless change the weights, and finite-step attacks can behave differently. For a fixed multiclass model with one shared scale, the smooth cap preserves the complete logit ordering, whereas hard clipping can add saturation ties but cannot reverse an ordering; Appendix A.3 states the exact assumptions.

All proofs concern real arithmetic. Smooth caps on 16-bit floating-point (FP16) or 16-bit brain floating-point (BF16) inputs use 32-bit floating-point (FP32) arithmetic internally and return the input data type. The execution checks use an outward representable magnitude bound and dtype-aware tolerances; they are kept separate from the exact mathematical statements (Appendix A.4).

4 Experimental design

The design asks whether a calibrated activation bound improves decisions, not merely whether it limits a tensor. We keep the training and primary threat model matched, retain every configuration, and distinguish implementation validity, intervention activity, and practical endpoint effects.

4.1 Matched training and distinct evaluation populations

We fine-tune the six-layer DistilRoBERTa classifier on IMDb sentiment classification (Maas et al., 2011). Its model family, RoBERTa (Robustly Optimized BERT Pretraining Approach), builds on Bidirectional Encoder Representations from Transformers (BERT) (Liu et al., 2019). The model has hidden size 768 and 12 attention heads; inputs are right-truncated to 256 tokens and dynamically

padded using the pinned model tokenizer. A deterministic stratified split assigns 22,500 labeled training reviews to fitting and 2,500 to validation. The 2,048-review activation-calibration subset remains part of the training population, not a separate test set.

The primary matrix comprises an uncapped control and hard/smooth caps at five locations, each trained with seeds $\{11, 23, 47, 89, 131\}$: 55 seed/configuration records, but only five matched model seeds. Within a seed, the initialization and training-order policy are shared. All parameters are fine-tuned for three epochs with Adam using decoupled weight decay (AdamW), learning rate 2×10^{-5}, weight decay 0.01, batch size 32, BF16, and gradient clipping at 1.0. Retaining the final incomplete minibatch gives 704 updates per epoch and 2,112 updates overall, with 212 warmup updates followed by linear decay. We select the epoch checkpoint with lowest validation negative log-likelihood (NLL), then higher BA for ties within 10^{-6}, then the earliest epoch. Every reported endpoint is linked to that selected checkpoint. Optimizer, sampling, and other details are in Appendix B.

Official clean evaluation uses all 25,000 test reviews. Embedding attacks use a fixed balanced subset of 1,000; stronger diagnostics use 200 source-matched reviews nested within it. Consequently, attack-induced drops use clean predictions on the same attacked subset, not the full-clean aggregate. Different selection salts produced different identifier namespaces; the recovered source crosswalk resolves the nesting. The 200-example comparison is a matched diagnostic, not independent replication. Test-access records and the post-access repair/exclusion chronology are in Appendix B.8.

4.2 Adaptive representation-space evaluation

For model seed k, let $e_t \in \mathbb{R}^{768}$ be the representation of an eligible token after embedding LayerNorm. The white-box attacker knows the weights, cap locations, and cap parameters, and perturbs this representation immediately before the optional embedding cap:

$$e'_t = e_t + \delta_t, \qquad \|\delta_t\|_2 \le \epsilon_k = \beta r_k, \qquad r_k = \mathrm{median}_{t \in \mathcal{C}_k} \|e_t^{(0,k)}\|_2. \tag{5}$$

Here $e_t^{(0,k)}$ denotes the matched control's representation, and $\mathcal{C}_k$ contains its non-special, non-padding training-calibration token occurrences. Special-token and padding perturbations are fixed to zero. For L eligible tokens, the product of tokenwise balls implies $\|\Delta\|_F \le \epsilon_k \sqrt{L}$; r_k is a token-level reference norm, not a normalization of the full sequence. The same r_k defines the feasible set for all configurations of seed k. This internal representation-space threat model does not assert that every feasible perturbation is a grammatical or meaning-preserving text edit.

The primary budget is $\beta = 0.05$; 0.025 and 0.10 are diagnostic budgets. Our projected gradient descent (PGD) attack uses ascent steps to maximize the untargeted final logit margin (Madry et al., 2018) $\max_{c \neq y} u_c - u_y$, where u denotes centered logits after any output cap. It uses 20 steps, three restarts, step size $2.5\epsilon_k/20$, random initialization in each token's ball, and per-token projection after every step. Each restart contributes its final iterate; the retained endpoint is the restart with the largest final margin. Intermediate trajectory maxima were not retained. Hard caps retain their exact forward map with an identity BPDA backward pass during attack optimization (Athalye et al., 2018); smooth caps use their true derivatives. For training, hard caps use the native clamp subgradient, not BPDA.

On the common 200 examples, 50-step, ten-restart attacks at $0.05r_k$ and $0.10r_k$ check optimization adequacy. A decrease exceeding two percentage points triggers the stronger-attack warning; native-hard versus BPDA is reported separately because its numerical materiality threshold was not prespecified. Constraint and finite-state checks accompany these comparisons. We report accuracy under the specified attack as robust balanced accuracy, not as certified worst-case accuracy. In particular, a small diagnostic drop near an accuracy floor is limited evidence of attack adequacy, not proof that a stronger adversary cannot exist.

4.3 Endpoints, uncertainty, and practical decisions

Primary endpoints are full-clean BA, primary-budget robust balanced accuracy (RBA), and *raw* attacked 15-bin expected calibration error (ECE-15) (Guo et al., 2017). NLL and the summed two-class Brier score support calibration interpretation; temperature scaling uses existing validation fits only as sensitivity analysis. For each seed, we orient cap/control effects so that positive always favors the cap:

$$\Delta_{\mathrm{clean}} = \mathrm{BA}_c - \mathrm{BA}_0, \quad \Delta_{\mathrm{robust}} = \mathrm{RBA}_c - \mathrm{RBA}_0, \quad \Delta_{\mathrm{cal}} = \mathrm{ECE}_0 - \mathrm{ECE}_c. \tag{6}$$

Subscripts c and 0 denote a capped configuration and its matched control.

Sample standard deviations (SDs) describe seed variation; confidence intervals (CIs) quantify uncertainty. The five paired seed differences, not examples or restarts, govern primary inference. Each operator/endpoint combination forms a separate five-location family: two-sided paired t tests use Holm correction at 0.05 (Holm, 1979). Individual 99% difference intervals provide at least 95% Bonferroni coverage within that family under the paired-t assumptions, not across all 30 contrasts. Practical equivalence uses two one-sided tests (TOST) within $(-0.01, 0.01)$ (Schuirmann, 1987; Lakens, 2017). For location j, we set $p_j^{\mathrm{eq}} = \max(p_{L,j}, p_{U,j})$ and Holm-adjust the five location-level values in each operator/endpoint family. Equivalence is not inferred from nonsignificance or from a difference interval. The primary margin is one percentage point (pp); a small detectable loss can still fall within it.

A robustness-benefit claim additionally requires a mean gain of at least 0.01, a corrected significant test, a favorable simultaneous interval, clean noninferiority, calibration non-worsening, verified engagement, and adequate adaptive evaluation. Direct location and hard/smooth claims have their own prespecified tests and practical rules (Appendix B.6). Historical activity, mechanism, and timing records keep their original measurement stage. QKNorm (five seeds) and FreeLB-style training (three seeds) are contextual comparators, not additions to the primary test families. These distinctions let Results address three separate questions: whether clean utility is retained, whether robustness improves practically, and what the measured budget regime reveals.

5 Results

All 55 selected primary checkpoints contribute. Training-only artifacts and execution checks identify the intervention; endpoint tests establish what it achieved. Historical mechanism records mark 45 of the 50 capped runs active and five smooth centered-logit runs inactive (Table 1). Keeping the latter distinguishes an enabled cap from a materially active intervention. Matched adaptive-attack diagnostics below then test the optimization used to evaluate these models.

5.1 Clean utility is preserved; practical robustness benefit is not established

The clean-performance question has a clear answer under the chosen policy: *all ten caps pass corrected clean-BA equivalence within ± 1 pp*. Table 1 reports the absolute results, while Figure 2(a) retains their paired differences and uncertainty. The control achieves 92.37% clean BA on 25,000 reviews and 1.56% robust BA on the separate 1,000-review primary attack subset. These scores describe different evaluation populations, not a matched clean-to-attacked drop.

Preserving clean utility does not yield the registered practical robustness benefit. In Figure 2(b), no cap meets the full benefit rule. Hard attention output has the largest mean gain, $+0.380$ pp, with the stored 99% difference interval $[-1.441, 2.201]$ pp. Its point estimate is below 1 pp, but the interval still permits useful positive effects.

The endpoint pattern is more informative than a single ranking (Table 1; Figure 2). Embedding caps are clean- and robust-equivalent under both operators; attacked-ECE equivalence holds for smooth but remains unresolved for hard embedding. Q/K and MLP caps preserve clean BA but leave robustness and attacked ECE unresolved for both operators. Smooth attention output is clean- and robust-equivalent but calibration-unresolved; both centered-logit variants pass all three tests. Historical smooth-logit inactivity makes its near-no-op result a control on interpretation, not evidence of an active defense. Among 27 contrasts for caps with recorded activity, 15 are equivalent and 12 unresolved; the inactive smooth centered-logit comparator adds 3 equivalence decisions. Overall, 18 of 30 contrasts are equivalent: 10 clean, 5 robust, and 3 attacked-ECE.

Equivalence does not mean exact equality. Hard MLP loses 0.658 pp clean BA and 0.880 pp robust BA, and worsens attacked ECE by 0.971 pp; its three difference intervals exclude zero (Figure 2). These detectable losses remain below the registered 1 pp requirement for practical harm. Its clean equivalence thus coexists with a small nonzero loss. E/U labels use Holm-adjusted location-level TOST p-values, not the plotted intervals. Panel (c) also shows why calibration needs its own endpoint: only three attacked-ECE contrasts are equivalent, and none hit the full calibration-benefit rule.

Table 1: Complete primary comparison on the official test. Entries are mean ± sample SD over five matched model seeds, not confidence intervals. Clean BA uses all 25,000 reviews; robust BA and raw ECE-15 use the 1,000-review primary attack at $0.05r$. All numbers use a percent scale; lower ECE is favorable. Rows follow the fixed intervention order, not performance. No configuration establishes the registered practical robustness-benefit rule.

Operator / site	Clean BA ↑	Robust BA ↑	ECE-15 ↓	Act.	C/R/E
Uncapped control	92.37 ± 0.22	1.56 ± 0.22	97.55 ± 0.34	—	—
Hard cap					
Embedding	92.38 ± 0.21	1.20 ± 0.40	97.87 ± 0.62	A	E/E/U
Q/K	92.16 ± 0.30	0.82 ± 0.52	98.52 ± 0.60	A	E/U/U
Attention output	92.32 ± 0.23	1.94 ± 0.94	96.98 ± 1.18	A	E/U/U
MLP	91.71 ± 0.20	0.68 ± 0.19	98.52 ± 0.34	A	E/U/U
Centered logits	92.35 ± 0.22	1.54 ± 0.27	97.54 ± 0.43	A	E/E/E
Smooth cap					
Embedding	92.29 ± 0.17	1.54 ± 0.40	97.48 ± 0.48	A	E/E/E
Q/K	92.36 ± 0.15	1.14 ± 0.36	97.84 ± 0.64	A	E/U/U
Attention output	92.31 ± 0.29	1.44 ± 0.57	97.30 ± 0.86	A	E/E/U
MLP	92.03 ± 0.30	0.76 ± 0.43	98.54 ± 0.65	A	E/U/U
Centered logits	92.37 ± 0.22	1.56 ± 0.22	97.55 ± 0.34	I	E/E/E

Act.: A/I = active/inactive in historical mechanism measurements, not newly measured official-test activity. **C/R/E**: corrected TOST decision for clean BA, robust BA, and raw attacked ECE; E = equivalent within the specified margin, U = unresolved practical difference.

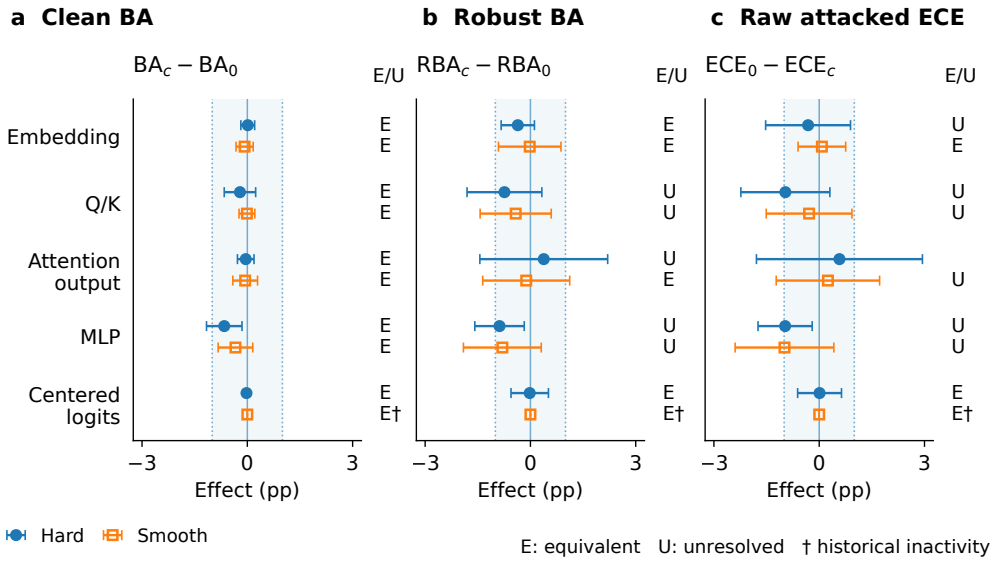

Figure 2: Effects, uncertainty, and equivalence are distinct. All contrasts use five paired seeds; positive favors the cap (method minus control for clean and robust BA, control minus method for raw attacked ECE). Bars are individual 99% difference intervals, giving 95% simultaneous coverage within each five-location operator/endpoint family, not across all 30 contrasts. Shading marks the ±1 pp reference margin. E/U uses Holm-adjusted location-level intersection–union TOST p-values, not interval appearance. A dagger marks historical smooth-logit inactivity.

Whether one location improves on another requires a direct comparison. The repeated-measures analyses detect location variation in all six operator/endpoint families, yet none of the 60 corrected location contrasts or 15 hard/smooth contrasts satisfies its complete practical-difference rule (Appendix C.2). Location can affect the measured outcome without identifying a better insertion point.

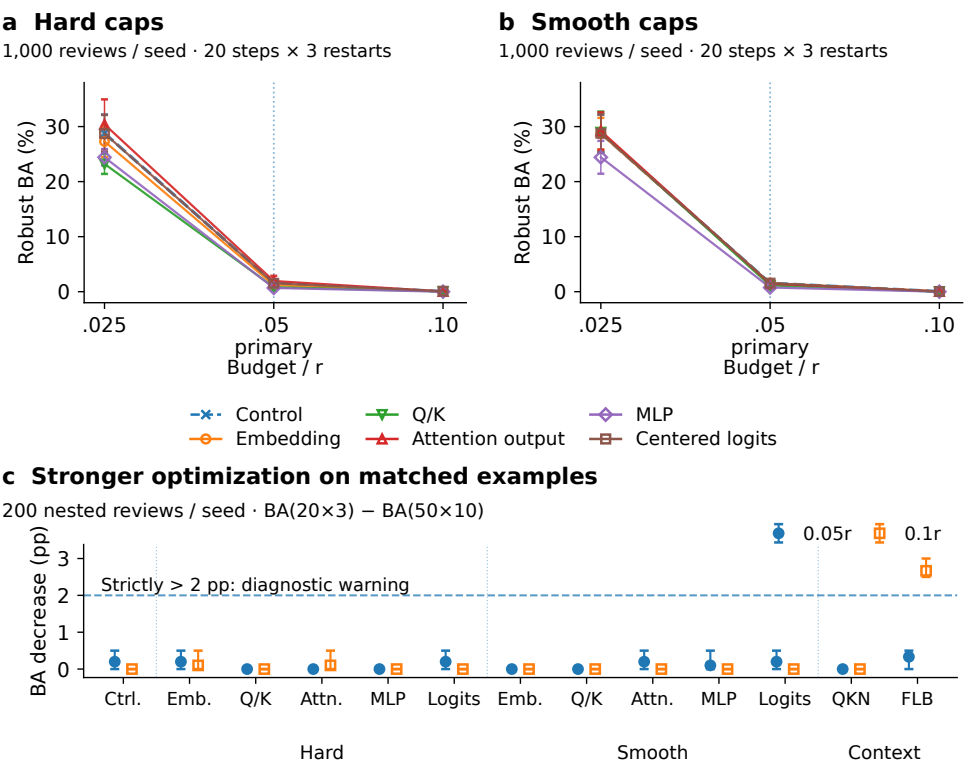

Figure 3: The measured threat regime and a separate matched diagnostic. (a,b) All eleven primary configurations at the three executed budgets; control is repeated to aid comparison. Points are five-seed means and bars are sample SDs, not confidence intervals. Budget spacing is logarithmic (doublings); lines only join measured points. Only $0.05r$ controls primary inference. (c) BA decrease from 20-step/3-restart to 50-step/10-restart optimization on the same 200 nested reviews per checkpoint. Points are stored means; whiskers are seed minima and maxima, not confidence intervals. Primary configurations and QKNorm (QKN) use five seeds; FreeLB-style (FLB) uses three. Its $0.10r$ drop exceeds 2 pp. Primary starting BA is at most 1%, so a drop strictly above 2 pp is impossible and the warning rule is non-diagnostic there. This panel is neither independent replication nor a native-versus-BPDA comparison.

5.2 Budget dependence and matched attacks qualify the primary result

Figure 3(a,b) retains all eleven configurations at all three budgets. A post-hoc analysis of the existing $0.025r$ results finds no corrected positive cap–control difference (Appendix Table 10). Hard Q/K is -5.56 pp lower, with a 99% interval of $[-10.34, -0.78]$ pp and Holm-adjusted $p = 0.029$; two centered-logit contrasts are equivalent and the other 7 unresolved. Hard attention output remains descriptively $+1.62$ pp but its interval spans zero. This sensitivity does not replace the registered $0.05r$ endpoint or its multi-endpoint benefit rule.

Panel (c) compares 20×3 with 50×10 optimization on the same 200 reviews. Every primary starting mean is at most 1%, so a decrease strictly above the prespecified 2 pp warning threshold is impossible; that rule cannot diagnose convergence for these rows (Appendix Table 11). The separate native-hard/BPDA comparison in Table 12 tests the backward route, not additional steps or restarts.

5.3 Context separates activation bounds from robust training

The FreeLB-style comparator (Zhu et al., 2020) supplies an informative contrast: its primary robust-BA effect is $+59.333$ pp against its three matched controls, with an ordinary paired 95% interval of $[58.329, 60.337]$ pp (Appendix Table 13). Thus the evaluator distinguishes at least one substantially more robust trained system. This is neither a compute-matched attribution nor a prespecified supe-

riority test against each cap. The qualification is budget-specific: its matched stronger-attack drop is 0.333 pp at $0.05r$, but 2.667 pp at $0.10r$, exceeding the 2 pp diagnostic criterion (Figure 3(c)). QKNorm's unfavorable contextual effects in Table 13 are likewise about this implementation and task, not normalization in general.

Supporting measurements sharpen this distinction. Stored-temperature sensitivity can reduce ECE without making the classifier robust (Figure 4). Greater historical tail compression tracks less favorable effects, while smooth Q/K and MLP caps incur throughput costs (Table 14; Figure 5); neither supplies a causal explanation or new test-set measurement. TextFooler outcomes and reviewer judgments separately qualify algorithmic vulnerability and textual validity (Tables 15 and 16). The primary finding remains: controlled magnitudes do not establish a practical robustness benefit.

6 Discussion and limitations

Magnitude control is a property, not an outcome. Distortion matching controls operator strength, not downstream robustness. Clean equivalence demonstrates utility preservation; Proposition 3 shows that output capping can alter confidence without changing fixed binary decisions. Training effects therefore require matched checkpoints and uncertainty.

The scope is deliberately exact. We test fixed, training-calibrated bounds in one binary DistilRoBERTa–IMDb classifier under matched seeds and attacks. With five paired seeds, a 99% interval has half-width about $2.06s_{\Delta}$; the primary floor further limits separation. At $0.025r$, no cap has a corrected positive effect, hard Q/K is 5.56 pp lower, and seven contrasts remain unresolved. FreeLB-style training identifies a substantially more robust regime, but its three-seed scope and stress-budget warning limit attribution. Learned or dynamic bounds remain as future work.

Supporting evidence tests alternative explanations. Historical compression associations remain exploratory because only five seeds underlie the rows. TextFooler ratings were collected independently under blinded conditions with consent, but remain conditional on successful attacks. Inactivity and timing retain their historical stages. The incompatible-tokenizer decoder evaluation is excluded from benchmark and replication claims; Appendix B.8 records that boundary. The classifier conclusion rests on independent evidence.

7 Conclusion

This study establishes a principled boundary between activation control and adversarial robustness. Distortion matching makes fixed hard and smooth caps comparable on training-only activations without equating their numerical scales. Across five transformer locations and five matched DistilRoBERTa seeds, all ten capped configurations preserve clean balanced accuracy within one percentage point, yet none establishes the prespecified primary robustness benefit. Among the 27 endpoint contrasts involving caps with recorded activity, 15 establish corrected equivalence and 12 remain unresolved; the inactive smooth centered-logit comparator contributes three additional equivalence decisions. The lower-budget analysis reaches the same central conclusion: no cap shows a corrected positive difference, while hard query/key capping is 5.56 percentage points worse than control. By contrast, FreeLB-style training occupies a substantially more robust regime under the same primary evaluation.

The broader lesson is that a bounded representation is not, by itself, a robust decision rule. Fixed caps can preserve clean utility, satisfy local mathematical bounds, and remain active within the network without producing a practically meaningful adversarial gain. Robustness claims therefore require direct decision-level evidence, adaptive evaluation, explicit intervention checks, and uncertainty analyses that distinguish equivalence from unresolved effects. These results provide both a comparison method and a clear evaluation standard for future activation-control studies. The empirical conclusion is specific to fixed, training-calibrated bounds in binary classification; learned, adaptive, and multiclass extensions should be judged by the same evidentiary standard.

ARTIFICIAL INTELLIGENCE (AI) USE STATEMENT

ChatGPT and Codex assisted with evidence organization, numerical consistency checks, prose drafting and revision, and table preparation; Codex also supported reproducibility checks. Scite assisted literature retrieval and bibliographic verification. The authors reviewed and verified all output and retain responsibility for the submission's claims, citations, mathematics, code, and reporting.

REPRODUCIBILITY STATEMENT

Section 3 and Appendix A provide the assumptions and complete proofs. Section 4 and Appendix B specify the data, revisions, calibration, training, attacks, and statistical decisions. Appendix C reports per-seed endpoints and supporting measurements. The supplementary data contain full-precision numerical results, per-seed measurements, attack budgets, and calibrated thresholds in comma-separated-value (CSV) files, together with evaluation-population definitions. Appendix E describes their contents. The release supports inspection of the reported findings; it does not include the training and attack implementations or an executable research environment.

ETHICS STATEMENT

This study evaluates adversarial vulnerability in a sentiment classifier; its findings neither certify the tested caps nor establish general language-model safety. Five reviewers consented to participate, served solely as reviewers, and submitted their ratings independently under blinded conditions. The reviewer-specific summaries characterize sampled successful attacks; Appendix E.3 reports their coverage and notes that exposure time was not recorded. Neither the paper nor the supplementary data contain raw review text, token arrays, individual rating forms, reviewer identities, model weights, or private operational endpoints. The invalid official language-model outputs are disclosed and excluded from benchmark and replication claims.

REFERENCES

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 274–283. PMLR, 2018. URL https://proceedings.mlr.press/v80/athalye18a.html.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. In *Advances in Neural Information Processing Systems*, volume 36, pp. 75067–75096, 2023. https://doi.org/10.52202/075280-3282.

Mohammed Bouri, Mohammed Erradi, and Adnane Saoud. ART: Attention-regularized transformers for multi-modal robustness. In *Findings of the Association for Computational Linguistics: EACL 2026*, pp. 5510–5535. Association for Computational Linguistics, 2026. https://doi.org/10.18653/v1/2026.findings-eacl.292.

Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. arXiv preprint arXiv:1902.06705, 2019.

Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic ReLU. In *Computer Vision – ECCV 2020*, volume 12364 of *Lecture Notes in Computer Science*, pp. 351–367. Springer, 2020. https://doi.org/10.1007/978-3-030-58529-7_21.

Jungwook Choi, Zhuo Wang, Swagath Venkataramani, Pierce I-Jen Chuang, Vijayalakshmi Srinivasan, and Kailash Gopalakrishnan. PACT: Parameterized clipping activation for quantized neural networks. arXiv preprint arXiv:1805.06085, 2018.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *Proceedings of the 37th International Conference on*

Machine Learning, volume 119 of *Proceedings of Machine Learning Research*, pp. 2206–2216. PMLR, 2020. URL https://proceedings.mlr.press/v119/croce20b.html.

Andrija Djurisic, Nebojsa Bozanic, Arjun Ashok, and Rosanne Liu. Extremely simple activation shaping for out-of-distribution detection. In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2209.09858.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Andrew F. Hayes and Klaus Krippendorff. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89, 2007. https://doi.org/10.1080/19312450709336664.

Alex Henry, Prudhvi Raj Dachapally, Shubham Shantaram Pawar, and Yuxuan Chen. Query-key normalization for transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 4246–4253. Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.findings-emnlp.379.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Zhichao Hou, Weizhi Gao, Yuchen Shen, Feiyi Wang, and Xiaorui Liu. ProTransformer: Robustify transformers via plug-and-play paradigm. In *Advances in Neural Information Processing Systems*, volume 37, pp. 137557–137609, 2024. https://doi.org/10.52202/079017-4370.

Yujia Huang, Huan Zhang, Yuanyuan Shi, J. Zico Kolter, and Anima Anandkumar. Training certifiably robust neural networks with efficient local lipschitz bounds. In *Advances in Neural Information Processing Systems*, volume 34, pp. 22745–22757, 2021. URL https://arxiv.org/abs/2111.01395.

Haoming Jiang, Pengcheng He, Weizhu Chen, Xiaodong Liu, Jianfeng Gao, and Tuo Zhao. SMART: Robust and efficient fine-tuning for pre-trained natural language models through principled regularized optimization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 2177–2190. Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.acl-main.197.

Di Jin, Zhijing Jin, Joey Tianyi Zhou, and Peter Szolovits. Is BERT really robust? a strong baseline for natural language attack on text classification and entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):8018–8025, 2020. https://doi.org/10.1609/aaai.v34i05.6311.

Sekitoshi Kanai, Masanori Yamada, Shin'ya Yamaguchi, Hiroshi Takahashi, and Yasutoshi Ida. Constraining logits by bounded function for adversarial robustness. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pp. 1–8. IEEE, 2021. https://doi.org/10.1109/IJCNN52387.2021.9533777.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses. *Social Psychological and Personality Science*, 8(4):355–362, 2017. https://doi.org/10.1177/1948550617697177.

Linyang Li, Ruotian Ma, Qipeng Guo, Xiangyang Xue, and Xipeng Qiu. BERT-ATTACK: Adversarial attack against BERT using BERT. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pp. 6193–6202. Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.emnlp-main.500.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

Andrew L. Maas, Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts. Learning word vectors for sentiment analysis. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pp. 142–150. Association for Computational Linguistics, 2011. URL https://aclanthology.org/P11-1015/.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=rJzIBfZAb.

Songping Peng, Zhiheng Zhang, Daojian Zeng, Lincheng Jiang, and Xieping Gao. Preventing safety drift in large language models via coupled weight and activation constraints. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 17646–17662. Association for Computational Linguistics, 2026. https://doi.org/10.18653/v1/2026.findings-acl.874.

Giovanni Puccetti, Anna Rogers, Aleksandr Drozd, and Felice Dell'Orletta. Outlier dimensions that disrupt transformers are driven by frequency. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1286–1304. Association for Computational Linguistics, 2022. https://doi.org/10.18653/v1/2022.findings-emnlp.93.

Vyas Raina, Samson Tan, Volkan Cevher, Aditya Rawal, Sheng Zha, and George Karypis. Extreme miscalibration and the illusion of adversarial robustness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2500–2525. Association for Computational Linguistics, 2024. https://doi.org/10.18653/v1/2024.acl-long.137.

Donald J. Schuirmann. A comparison of the two one-sided tests procedure and the power approach for assessing the equivalence of average bioavailability. *Journal of Pharmacokinetics and Biopharmaceutics*, 15(6):657–680, 1987. https://doi.org/10.1007/BF01068419.

Korn Sooksatra and Pablo Rivas. Dynamic-max-value ReLU functions for adversarially robust machine learning models. *Mathematics*, 12(22):3551, 2024. https://doi.org/10.3390/math12223551.

Korn Sooksatra, Gissella Bejarano, and Pablo Rivas. Evaluating robustness of reconstruction models with adversarial networks. *Procedia Computer Science*, 222:353–366, 2023. https://doi.org/10.1016/j.procs.2023.08.174.

Korn Sooksatra, Greg Hamerly, and Pablo Rivas. Is ReLU adversarially robust? arXiv preprint arXiv:2405.03777, 2024a.

Korn Sooksatra, Bikram Khanal, and Pablo Rivas. On adversarial examples for text classification by perturbing latent representations. arXiv preprint arXiv:2405.03789, 2024b.

David Stutz, Matthias Hein, and Bernt Schiele. Confidence-calibrated adversarial training: Generalizing to unseen attacks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9155–9166. PMLR, 2020. URL https://proceedings.mlr.press/v119/stutz20a.html.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=F7aAhfitX6.

Yiyou Sun, Chuan Guo, and Yixuan Li. ReAct: Out-of-distribution detection with rectified activations. In *Advances in Neural Information Processing Systems*, volume 34, pp. 144–157, 2021. URL https://arxiv.org/abs/2111.12797.

Boxin Wang, Shuohang Wang, Yu Cheng, Zhe Gan, Ruoxi Jia, Bo Li, and Jingjing Liu. InfoBERT: Improving robustness of language models from an information theoretic perspective. In *International Conference on Learning Representations*, 2021. https://doi.org/10.48550/arXiv.2010.02329. URL https://arxiv.org/abs/2010.02329.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 38087–38099. PMLR, 2023. URL https://proceedings.mlr.press/v202/xiao23c.html.

Jiehang Zeng, Jianhan Xu, Xiaoqing Zheng, and Xuanjing Huang. Certified robustness to text adversarial attacks by randomized [MASK]. *Computational Linguistics*, 49(2):395–427, 2023. https://doi.org/10.1162/coli_a_00476.

Huichi Zhou, Zhaoyang Wang, Hongtao Wang, Dongping Chen, Wenhan Mu, and Fangyuan Zhang. Evaluating the validity of word-level adversarial attacks with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 4902–4922. Association for Computational Linguistics, 2024. https://doi.org/10.18653/v1/2024.findings-acl.292.

Chen Zhu, Yu Cheng, Zhe Gan, Siqi Sun, Tom Goldstein, and Jingjing Liu. FreeLB: Enhanced adversarial training for natural language understanding. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=BygzbyHFvB.

Jiachen Zhu, Xinlei Chen, Kaiming He, Yann LeCun, and Zhuang Liu. Transformers without normalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14901–14911. IEEE/CVF, 2025. https://doi.org/10.1109/CVPR52734.2025.01388.

A COMPLETE MATHEMATICAL DETAILS

The finite-sample theorem justifies the operator comparison. The propositions below then identify exactly what a cap controls: local distances, score ranges, and confidence. None assumes that these properties by themselves guarantee correct adversarial decisions. All statements in this section use real arithmetic.

A.1 WHY THE MATCHING SCALE EXISTS AND IS UNIQUE

For Theorem 1, write $b_i = |a_i|$ and $\mu = N^{-1}\sum_i b_i$. It is important that μ is a mean *magnitude*, not the signed mean of the activations. The sample may contain negative values, zeros, repetitions, or just one nonzero value; no distributional assumption is needed. The three steps below expand the complete main-text proof.

Step 1: remove the absolute value without changing the objective. For $u > 0$, the identity $\tanh u = \int_0^u \operatorname{sech}^2 v\, dv$ implies $0 < \tanh u < u$, because $0 < \operatorname{sech}^2 v < 1$ for $v > 0$. Therefore the smooth cap has the same sign as a nonzero input and strictly smaller magnitude. Using oddness gives

$$|a_i - s\tanh(a_i/s)| = b_i - s\tanh(b_i/s). \tag{7}$$

For $b_i = 0$ both sides are zero. This identity, rather than a differentiation through an unspecified absolute-value branch, is the basis of the proof.

Step 2: show strict decrease with respect to scale. For a fixed magnitude $a > 0$, set $f_a(s) = a - s\tanh(a/s)$ and $u = a/s$. Differentiating with respect to s yields

$$\begin{aligned} f_a'(s) &= -\tanh u + u\operatorname{sech}^2 u = -g(u), \\ g(u) &= \tanh u - u\operatorname{sech}^2 u, \qquad g'(u) = 2u\operatorname{sech}^2 u\tanh u > 0 \quad (u > 0). \end{aligned} \tag{8}$$

Since $g(0) = 0$, we have $g(u) > 0$ on $(0, \infty)$ and hence $f_a'(s) < 0$. Zero sample terms are constant, but at least one nonzero term is present. Their finite average $D_S(s)$ is consequently continuous and strictly decreasing. The direction is intuitive: increasing s brings the smooth cap closer to the identity and reduces the amount of intervention.

Step 3: identify the attainable targets. As $s \downarrow 0$, $0 \le s\tanh(a/s) \le s$, giving $f_a(s) \to a$. For $a > 0$, as $s \to \infty$,

$$s\tanh(a/s) = a\frac{\tanh(a/s)}{a/s} \longrightarrow a, \tag{9}$$

because $\tanh u/u \to 1$ at zero. The corresponding zero term remains zero. A finite sum therefore gives the limits μ and 0 for D_S. For a target $d \in (0, \mu)$, choose finite positive $s_L < s_U$ far enough toward these endpoints that $D_S(s_L) > d > D_S(s_U)$. The intermediate value theorem on $[s_L, s_U]$ gives a solution; strict decrease forbids a second solution. This completes the existence and uniqueness argument.

Degenerate targets. For a nonzero sample and finite $s > 0$, every nonzero term lies strictly between zero and its input magnitude. Thus $0 < D_S(s) < \mu$: target zero is reached only as $s \to \infty$, and target μ only as $s \downarrow 0$. Targets outside $[0, \mu]$ are impossible. For an all-zero sample, D_S is identically zero, so every positive scale matches zero and uniqueness fails. These are mathematical endpoint facts, not reasons to silently substitute a convenient finite scale.

Why the hard target is admissible. For any $\tau > 0$ and $a > 0$, $(a-\tau)_+ < a$. Therefore a nonzero sample always has $D_H(\tau) < \mu$. The sole remaining condition is $D_H(\tau) > 0$, equivalently that at least one sampled magnitude strictly exceeds τ. If the threshold is at or above the largest magnitude, the hard cap is an exact no-op on the sample and no finite smooth scale can exactly match it.

Why the scales are unequal. For $a > 0$, both $\tanh(a/\tau) < a/\tau$ and $\tanh(a/\tau) < 1$, so

$$\tau\tanh(a/\tau) < \min(a, \tau). \tag{10}$$

At equal parameters, the smooth cap therefore changes every nonzero calibration value more than the hard cap. Averaging gives $D_S(\tau) > D_H(\tau)$. When the hard target is positive, the unique smooth match must consequently satisfy $s^\star > \tau$. This is an explanatory consequence of the stated operators, not a claim that smooth caps are stronger after their distortions have been matched.

Large-scale asymptotics and shallow matching. For the fixed finite sample, define $m_3 = N^{-1}\sum_i b_i^3$ and $m_5 = N^{-1}\sum_i b_i^5$. The Taylor expansion $\tanh u = u - u^3/3 + 2u^5/15 + O(u^7)$ is uniform over the finitely many values $u = b_i/s$ as $s \to \infty$. Substitution into Equation (7) gives

$$D_S(s) = \frac{m_3}{3s^2} - \frac{2m_5}{15s^4} + O(s^{-6}),$$
$$D_S'(s) = -\frac{2m_3}{3s^3} + \frac{8m_5}{15s^5} + O(s^{-7}). \tag{11}$$

Because a nonzero sample has $m_3 > 0$, the unique solution for a target $d \downarrow 0$ satisfies $s^\star(d) \to \infty$ and

$$s^\star(d) \sim \sqrt{\frac{m_3}{3d}}. \tag{12}$$

Thus tiny absolute matching targets can require large scales and can make the distortion objective shallow in s. The centered-logit artifacts illustrate this regime: their displayed hard thresholds range from 2.6875 to 3.1875, while the matched smooth scales range from 120.3623 to 216.4342 (Tables 8–9). This asymptotic explanation does not prove finite-precision identity or inactivity; those are empirical execution properties recorded separately.

Bisection and the recorded tolerance. A finite bracket $[s_L, s_U]$ with $D_S(s_L) \geq d \geq D_S(s_U)$ contains the unique root. At each midpoint, $D_S(s_{\mathrm{mid}}) > d$ moves the lower endpoint upward; $D_S(s_{\mathrm{mid}}) < d$ moves the upper endpoint downward. Equality identifies the root immediately. After j halvings the bracket width is $(s_U - s_L)/2^j$, and its midpoint is within half that width of the exact root. This proves convergence in scale, not a uniform numerical conditioning guarantee. The recorded artifact check is instead the distortion-residual criterion

$$\frac{|D_S(\widehat{s}) - D_H(\tau)|}{\max\{D_H(\tau), 10^{-12}\}} \leq 10^{-6}, \qquad D_H(\tau) > 0. \tag{13}$$

The 10^{-12} denominator floor stabilizes the reported residual; it does not make a zero-distortion artifact valid. The theorem concerns the real objective. A quantized program can have plateaus, and its stored scale and residual require separate numerical verification.

A.2 Local bounds and concentration

The following proposition makes the local consequences used by the execution checks explicit. It concerns ordinary scaled dot-product attention without an additional unbounded score bias, and attention probabilities before dropout.

Proposition 2 (Local bounds and deterministic concentration)**.** *Let T_b be C_b or S_b for $b > 0$.*

1. *For $x, y \in \mathbb{R}^m$ and $1 \leq p \leq \infty$, $\|T_b(x)\|_\infty \leq b$ and $\|T_b(x) - T_b(y)\|_p \leq \|x - y\|_p$.*
2. *Let $q, k_j \in \mathbb{R}^{d_h}$ satisfy $\|q\|_\infty, \|k_j\|_\infty \leq b$ for each of $n \geq 1$ allowed keys. For $\ell_j = q^\top k_j/\sqrt{d_h}$, set $B = \sqrt{d_h}b^2$ and $p_j = \exp(\ell_j)/\sum_{i=1}^n \exp(\ell_i)$. Then*

$$|\ell_j| \leq B, \qquad p_{\max} \leq \frac{1}{1 + (n-1)e^{-2B}}, \qquad H(p) \geq \log\bigl(1 + (n-1)e^{-2B}\bigr), \tag{14}$$

   where $H(p) = -\sum_j p_j \log p_j$ is Shannon entropy in nats.

3. *For $z \in \mathbb{R}^K$, $K \geq 2$, and $c \in \mathbb{R}$, $\mathrm{ctr}(z + c\mathbf{1}) = \mathrm{ctr}(z)$. If $u = T_b(\mathrm{ctr}(z))$ and $\pi = \mathrm{softmax}(u)$, then*

$$\max_j \pi_j \leq \frac{1}{1 + (K-1)e^{-2b}}. \tag{15}$$

Proof. The hard cap is the continuous, piecewise-linear scalar projection onto $[-b, b]$; its slopes are zero or one wherever the derivative exists. Integrating over any interval, including intervals crossing a breakpoint, gives $|C_b(x) - C_b(y)| \le |x - y|$. For the smooth cap, $S_b'(x) = \operatorname{sech}^2(x/b) \in (0, 1]$, and the mean value theorem gives the same scalar inequality. Applying it to each coordinate and then taking an ℓ_p norm proves nonexpansiveness, including the maximum norm at $p = \infty$. The definitions directly give the magnitude bounds.

For each allowed attention score, $|q^\top k_j| \le \sum_{r=1}^{d_h} |q_r||k_{jr}| \le d_h b^2$, which proves $|\ell_j| \le B$. Let $j_\star$ attain the largest score. Since every score difference is at most $2B$,

$$p_{j_\star} = \frac{1}{1 + \sum_{i \ne j_\star} \exp(\ell_i - \ell_{j_\star})} \le \frac{1}{1 + (n-1)e^{-2B}}. \tag{16}$$

Also $-\log p_j \ge -\log p_{\max}$ for every j. Multiplying by p_j and summing gives $H(p) \ge -\log p_{\max}$ and hence the entropy bound. For $n = 1$, the bounds correctly reduce to probability one and entropy zero.

Finally, $K^{-1}\mathbf{1}^\top(z + c\mathbf{1}) = K^{-1}\mathbf{1}^\top z + c$, so centering removes the additive constant exactly. Every coordinate of u lies in $[-b, b]$, so any pairwise logit difference is at most $2b$. Applying the same denominator argument over K classes gives Equation (15). □

Masks, residuals, and scope. Only allowed keys enter Equation (14). A padding mask is applied after finite scores are formed, and a masked probability is zero. An all-masked row has no probability distribution of this form and is outside the assumption $n \ge 1$; a numerical convention for such a row is a separate kernel contract, not an entropy theorem. The entropy bound is conservative, not a claim of a tight minimum. Nor does the proposition bound the residual stream merely because one branch is capped, or assert that subsequent projections and normalizations preserve the cap's local Lipschitz constant. Centering is performed *before* the output cap; for $K > 2$ a coordinatewise cap need not retain a zero-sum output vector.

Numerical scale of the attention bound. The classifier has $d_h = 64$ and at most $n = 256$ allowed token positions. The median of the 30 displayed hard Q/K thresholds is $b = 6.0$, which gives $B = \sqrt{64}b^2 = 288$ and $(n-1)e^{-2B} = 255e^{-576} \approx 1.79 \times 10^{-248}$. Hence the probability ceiling in Equation (14) is $1 - 1.79 \times 10^{-248} + O(10^{-496})$, while the entropy floor is approximately 1.79×10^{-248} nats. The displayed smooth Q/K scales are larger (16.7796–26.9252), making the same worst-case bounds still looser. The theorem is correct, but at these calibrated scales it supplies essentially no practical anti-concentration guarantee. This numerical context neither shows that attention actually concentrated nor explains the empirical outcomes causally.

A.3 Output-only capping and fixed binary decisions

Proposition 3 (Output magnitude and binary decisions)**.** *Let $z(x) \in \mathbb{R}^2$ be the logits of a fixed classifier and let h be C_b or S_b with $b > 0$. In real arithmetic, applying h coordinatewise to $\operatorname{ctr}(z(x))$ preserves the argmax set at every input. Under the same tie-breaking rule, it therefore preserves the decision function.*

Proof. Write $d = z_1(x) - z_2(x)$. The centered logits are $(d/2, -d/2)$, and their capped difference is

$$h(d/2) - h(-d/2) = 2h(d/2). \tag{17}$$

Both maps are odd and preserve the sign of every nonzero input. The right-hand side is positive when $d > 0$, negative when $d < 0$, and zero exactly when $d = 0$. Thus each unique maximizer, and the two-class tie, are preserved. □

This is an elementary explanatory observation, not a claimed algorithmic advance. For the same fixed model and the same admissible perturbation set, the exact sets of correctly and incorrectly classified inputs are unchanged by output-only capping. Finite-step attack optimization can still return different inputs because its objective and gradients change. Training with a cap can change the model itself.

Multiclass ordering. Mean centering subtracts the same scalar from every logit and therefore preserves all coordinate orderings for any K. With one shared positive scale, S_b is strictly increasing because $S_b'(x) = \operatorname{sech}^2(x/b) > 0$; smooth output capping therefore preserves the full ordering and argmax set of a fixed multiclass model in real arithmetic. Hard clipping C_b is nondecreasing, so it cannot reverse an ordering, but its flat saturation regions can turn strict inequalities into ties. A deterministic tie-breaking rule can then select a different class only through such a newly introduced tie. These statements do not cover class-dependent scales, finite-precision loss of strict order, or models retrained with the cap.

A.4 Finite-precision execution is a separate statement

For FP16 and BF16 inputs, the implemented smooth cap evaluates division, tanh, and multiplication internally in FP32 and performs one final conversion back to the input dtype. FP32 and 64-bit floating-point (FP64) inputs retain their native arithmetic path. The nominal calibrated scale is not retuned to eliminate a numerical discrepancy. Writing R_d for output conversion and fl_{32} for the internal floating-point computation, the low-precision operation is more accurately described as

$$\widehat{S}_{s,d}(x) = R_d(\mathrm{fl}_{32}[s\tanh(x/s)])\,, \qquad b_{\mathrm{exec}}(s,d) = \min\{v \in \mathbb{F}_d : v \geq s\}, \tag{18}$$

where $\mathbb{F}_d$ is the set of finite representable numbers of dtype d and the bound is defined when that set contains a value at least s.

If the inner real expression were evaluated exactly, monotone rounding would bound its returned magnitude by $b_{\mathrm{exec}}(s,d)$. The actual computation also rounds internally, however. Formula agreement, returned-tensor range, sign, finite-state, and dtype-specific contraction checks therefore remain separate execution obligations. Likewise, score and probability calculations require their recorded numerical tolerances; substituting an outward bound does not turn an entire floating-point attention pipeline into an exact-arithmetic proof. The recorded FP32 embedding observations include expansions of one unit in the last place (ULP), whereas the BF16 artifact checks address the quantized path. Neither is a refutation of Theorem 1, and neither justifies silently weakening or relabeling an executed numerical check. These distinctions preserve the real theorem and the empirical verification as different kinds of evidence.

B PROTOCOL DETAILS AND INTERPRETATION BOUNDARIES

This appendix makes the intervention and evaluation reconstructible without conflating design specifications, measured outcomes, and later source verification. The primary classifier contracts, corrected source-overlap records, and official analysis define the scientific statements below. Historical validation results do not replace official-test results.

B.1 INSERTION SITES AND PARAMETER SHARING

Table 2 specifies the tensors being capped. The distinction between a branch output and the residual stream is consequential: capping an attention branch does not cap the residual stream after addition. Similarly, the MLP intervention precedes its output projection and therefore does not bound the projected branch by the same scalar threshold. All five configurations retain the underlying model's normalization, residual, dropout, and masking operations.

Table 2: Exact cap locations in the classifier. One row is selected per capped configuration. The three block-local rows apply in every layer, with one bound per layer; embedding and centered-logit bounds are global for their respective sites. Q and K share their layer's calibrated bound. Masking sentinels are never capped.

Identifier	Tensor and operation order	Bounds per seed
`emb_post_ln`	Normalized embeddings, after embedding LayerNorm and before embedding dropout.	1
`qk_post_proj`	Q and K after linear projection and head reshape; before their scaled dot product and the mask.	6; Q/K pooled
`attn_out_pre_dropout`	Concatenated-head output after its projection; before branch dropout, residual addition, and LayerNorm.	6
`mlp_post_gelu`	Feed-forward activations after GELU; before the output projection.	6
`centered_logits`	Per-example class logits after mean centering; before cross-entropy, softmax, and attack scoring.	1

The resulting 20 calibration identities are one embedding, six pooled Q/K, six attention-output, six MLP, and one centered-logit identity. Each stores a hard threshold and its matched smooth scale; it is not two independent calibration samples. Across five seeds there are 100 such artifacts. A threshold is fixed across token positions, coordinates, and attention heads within its calibrated site/layer. The cap is the identity when disabled. The centered-logit operation is shared by the control and all capped configurations.

B.2 TRAINING-ONLY CALIBRATION AND ITS EXACT MEANING

Calibration uses the selected uncapped checkpoint with dropout disabled and no gradient calculation. The 2,048 reviews are balanced by class and remain members of the 22,500-review fitting set. They are not used as a held-out estimate of performance. Padding coordinates are excluded from activation sampling. For each example, the documented sampler selects without replacement 64 embedding values, 32 Q and 32 K values per layer, 64 attention-output values per layer, and 64 post-GELU values per layer. Thus each non-logit identity contributes 64 candidates per review, or 131,072 candidates over 2,048 reviews; Q and K are pooled within a layer. The execution contract then applies the protocol's deterministic bounded capacity of 65,536 values per non-logit identity. Both centered classifier logits are retained for every review, giving 4,096 values. The accepted threshold artifacts bind these final sample counts and sample hashes. Their evidence contract does not encode the internal coordinate ordering used to enforce the bounded capacity, so we report the verified candidate and retained counts without claiming an unrecorded ordering rule.

The primary quantile is fixed at $q = 0.995$. If $b_{(1)} \leq \cdots \leq b_{(N)}$ are the sorted sampled magnitudes, the frozen mathematical specification writes

$$\tau = b_{(\lceil qN \rceil)}. \tag{19}$$

Here `higher` names that documented convention, not an assumed software library default. The expressions $\lceil qN \rceil$ and $1 + \lceil q(N-1) \rceil$ agree at the two recorded sample sizes: rank 65,209 for

$N = 65,536$ and rank 4,076 for $N = 4,096$. The released numerical records bind the thresholds, ranks, and matching residuals but do not include the original helper implementation. Equation (19) therefore states the documented convention; both historical rank expressions agree at the recorded sample sizes.

The matching residual is defined in Equation (13). Accepted artifacts require finite samples, a positive finite threshold, positive hard distortion, a positive finite smooth scale, and matching residual at most 10^{-6}. They also require hard exceedance between 0.0025 and 0.0075 and matching checkpoint, calibration-manifest, sample-count, and parameter identities. Exceedance means $|a_i| > \tau$, including for the smooth-cap threshold-exceedance diagnostic; it is not the same as the fraction changed by a smooth function. Once accepted, both operators' parameters remain fixed during capped training. A low or zero observed engagement state is retained, not used to retune one model.

The versioned random keys use compact canonical JavaScript Object Notation (JSON) arrays, Unicode Normalization Form C (NFC), the 8-bit Unicode Transformation Format (UTF-8), and fixed field order. Per-example coordinate sampling and attack randomness have separate namespaces. The supplementary numerical records retain the resulting identifiers and hashes, preserving provenance without reconstructing identities from informal text concatenation.

B.3 Model, training, and checkpoint selection

Table 3 records the primary training policy. We use the byte-level byte-pair encoding (BPE) tokenizer corresponding to the pinned DistilRoBERTa snapshot. The exact revisions are:

Model: `fb53ab8802853c8e4fbdbcd0529f21fc6f459b2b`
IMDb: `e6281661ce1c48d982bc483cf8a173c1bbeb5d31`

The model identifier is `distilbert/distilroberta-base`; the labels are $0 =$ negative and $1 =$ positive. The standard two-label classification head consumes the first-token hidden state.

Table 3: Matched primary-classifier training policy. Capped models start from the pretrained initialization, not from the trained control used for calibration. The final partial minibatch is retained, so the step counts are exact.

Item	Setting
Model	Six layers; hidden size 768; 12 heads; head dimension 64
Inputs	Maximum 256 tokens; right truncation; dynamic padding
Training / validation	22,500 / 2,500 labeled reviews, stratified by class
Seeds	11, 23, 47, 89, 131
Optimizer	AdamW; learning rate 2×10^{-5}; weight decay 0.01
Optimizer details	Betas $(0.9, 0.999)$; epsilon 10^{-8}; gradient clipping 1.0
Batch policy	Train 32; evaluate 64; accumulation 1; no data-loader workers
Schedule	Three epochs; 704 steps/epoch; 2,112 total; 212 warmup steps, linear decay
Selection	Validate at steps 704, 1,408, 2,112; lowest NLL, then highest BA within 10^{-6}, then earliest epoch
Execution	BF16; eager attention; no compilation; TensorFloat-32 (TF32) disabled; deterministic algorithms requested
Other training choices	Full-parameter fine-tuning; checkpoint-default dropout; no label smoothing, augmentation, or early stopping

The cap parameters are not trainable. Hard-cap training uses the native clamp subgradient, including the framework's breakpoint convention; smooth-cap training uses its derivative. The attack-only identity backward approximation is not a training modification. Selected-checkpoint, configuration, data-manifest, artifact, and execution identities accompany the measurements. The calibration activations that define the cap belong to the selected *control*; later endpoint measurements belong to the selected checkpoint of the particular capped run.

B.4 Evaluation populations and attack diagnostics

Table 4 separates populations that must not share a denominator by implication. The deterministic primary and strong subsets are shared across the relevant checkpoints. Their different namespace identifiers required source-level matching, which recovered all 200 strong examples inside the 1,000-example primary subset. That correction supersedes older disjoint-subset wording without turning the diagnostic into independent replication.

Table 4: Populations used by the classifier study and its supporting measurements. Counts are examples per checkpoint except for the reviewer sample. The full-clean endpoint and attacked endpoint use different populations. Historical measurements are not relabeled as official-test measurements.

Population	Count	Role and relationship
Official clean test	25,000	12,500 per class; full-clean endpoints
Primary attack subset	1,000	500 per class; budgets $0.025r$, $0.05r$, $0.10r$
Matched stronger-attack subset	200	100 per class; nested in the primary set; $0.05r$, $0.10r$
Historical mechanism subset	256	Training-derived measurements at selected checkpoints
TextFooler subset	250	125 per class; control and two validation-selected cap families
Human-rating sample	150 pairs	Selected from successful attacks; 111 distinct source reviews

Embedding PGD constrains each eligible token independently in its ℓ_2 ball, rather than constraining one aggregate norm over the sequence. If L tokens are eligible, then $\|\Delta\|_F^2 = \sum_{t=1}^{L} \|\delta_t\|_2^2 \leq L\epsilon^2$ and hence $\|\Delta\|_F \leq \epsilon\sqrt{L}$. Because r_k is a token-level median norm, this inequality is not a percentage change of the full sequence representation. The projected update cannot alter padding or special tokens. Every restart begins with a random feasible perturbation; the shared randomness is keyed by model seed, example identity, and restart. The primary optimizer uses 20 steps and three restarts at each reported budget. The stronger optimizer uses 50 steps and ten restarts, with step sizes $2.5\epsilon/20$ and $2.5\epsilon/50$, respectively. The chosen endpoint maximizes the *final* untargeted logit margin over restarts, rather than retaining the best candidate over all iterations. The accepted records retain the chosen restart, final attacked logits, and final margin, but no per-step trajectory maxima; a trajectory-best candidate cannot be reconstructed after the fact. The stronger-budget checks compare the same final-endpoint rule with more steps and restarts and do not turn it into trajectory-best optimization. This finite search does not exhaust the feasible set.

Hard-cap evaluation uses exact clamping in the forward pass and the identity approximation in the backward pass. Native-hard comparisons are recorded separately on the strong subset. This addresses a specific gradient-masking concern (Athalye et al., 2018; Carlini et al., 2019); it does not establish that every possible adaptive attack has been exhausted. A native/BPDA materiality threshold was not numerically fixed; the two-percentage-point rule applies only to the stronger-optimization comparison. Mode restoration, parameter-gradient preservation, finite outputs, and perturbation constraints are execution checks, not substitutes for adversarial success.

B.5 Endpoint definitions

For binary labels, let true positives (TP), false negatives (FN), true negatives (TN), and false positives (FP) count predictions against the labels. Balanced accuracy is

$$\mathrm{BA} = \frac{1}{2}\left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right). \tag{20}$$

Both required classes are present in the reported full evaluation subsets. Robust BA uses the predictions returned by the specified attack on those subsets. Attack success among clean-correct examples has its own denominator and is not $1 - \mathrm{BA}$ in general.

Let π_i be the class-probability vector for example i, y_i its label, $\widehat{y}_i$ its prediction, and $c_i = \max_j \pi_{ij}$ its confidence. For M evaluated examples and bins $B_1, \ldots, B_{15}$, ECE is

$$\mathrm{ECE}_{15} = \sum_{m:B_m \neq \varnothing} \frac{|B_m|}{M} \left| \frac{1}{|B_m|} \sum_{i \in B_m} \mathbf{1}\{\widehat{y}_i = y_i\} - \frac{1}{|B_m|} \sum_{i \in B_m} c_i \right|. \tag{21}$$

The fixed bins are $[(m-1)/15, m/15)$ for $m < 15$, with $[14/15, 1]$ as the final bin; confidence exactly one is included. Empty bins contribute zero. NLL is $-M^{-1}\sum_i \log \pi_{i,y_i}$ in nats. The implemented Brier score is $M^{-1}\sum_i \sum_{j\in\{0,1\}}(\pi_{ij} - \mathbf{1}\{y_i = j\})^2$, with range $[0, 2]$; it is not divided by the number of classes. Raw attacked probabilities define the primary calibration endpoint. Stored temperatures fitted on validation NLL, within $[0.05, 10]$, provide a separate sensitivity analysis, not a refit on official test outcomes.

B.6 Matched-seed statistical decisions

The three favorable cap/control differences are defined in Equation (6). Each configuration supplies five paired seed differences $\Delta_1, \ldots, \Delta_5$ with sample mean $\overline{\Delta}$ and sample standard deviation s_Δ. The primary difference interval is

$$\overline{\Delta} \pm t_{4,0.995}\frac{s_\Delta}{\sqrt{5}}, \tag{22}$$

where $t_{\nu,q}$ is the q quantile of the Student t distribution with ν degrees of freedom. This is a 99% *individual* interval and a Bonferroni construction with at least 95% simultaneous coverage, under the paired-t assumptions, for the five locations within an operator/endpoint family. There are six such families. Neither these intervals nor the corresponding within-family Holm corrections are described as controlling one family consisting of all 30 primary contrasts. With five paired seeds, $t_{4,0.995}/\sqrt{5} \approx 2.059$, so the interval half-width is about $2.06 s_\Delta$. For hard attention output at the primary budget, $s_\Delta = 0.884$ pp and the half-width is 1.821 pp, producing the reported $[-1.441, 2.201]$ pp interval around the $+0.380$ pp mean. This is a precision calculation, not a post-hoc power claim: power would additionally require an assumed population variance, alternative, test level, and target power. Five seeds also limit assessment of the paired-t assumptions; the recorded sensitivity analyses do not override the primary rule.

For equivalence margin $\eta = 0.01$, TOST tests the two nulls

$$H_{0,L} : \mathbb{E}[\Delta] \le -\eta, \qquad H_{0,U} : \mathbb{E}[\Delta] \ge \eta. \tag{23}$$

For location j, the two one-sided tests produce $p_{L,j}$ and $p_{U,j}$. Because equivalence rejects the union of the two directional nulls, its location-level intersection–union p-value is $p_j^{\mathrm{eq}} = \max(p_{L,j}, p_{U,j})$. Holm adjustment is applied once to the five p_j^{eq} values within each operator/endpoint family, and equivalence requires the adjusted value below 0.05. An ordinary 90% TOST interval is not substituted for this corrected decision, and the 99% difference interval is a different inferential object. A robustness benefit requires mean gain at least η, a corrected significant robust-BA difference, and an interval excluding zero in the favorable direction, plus the clean, calibration, engagement, and attack checks. The rule does not require the entire difference interval to exceed η. Clean noninferiority and calibration non-worsening each test a favorable difference greater than -0.01, with their specified one-sided within-operator corrections. A calibration-benefit statement additionally requires favorable NLL or Brier.

The already-executed $0.025r$ robust-BA results are additionally analyzed as an explicitly post-hoc sensitivity check motivated by the primary floor. The selector is fixed to the primary 1,000-example population, raw probabilities, 20 steps, three restarts, and the original five checkpoints per configuration. For each cap–control contrast we report the five paired differences, mean, paired SD, individual 99% paired-t interval, a Holm-adjusted two-sided difference test, and the corrected location-level equivalence p-value. Each Holm correction spans the five locations within one operator family; no simultaneous claim combines the two operators. The ± 1 pp margin is unchanged. This robust-BA-only sensitivity does not replace the prespecified $0.05r$ endpoint or satisfy the full clean/robustness/calibration/engagement benefit rule.

Location claims require a repeated-measures global test with seed as the block and the stated Greenhouse–Geisser correction. A significant global test permits ten paired location contrasts within that operator/endpoint family. Those use Holm correction and individual 99.5% intervals, with a point difference at least 0.01 required for a practical direct claim. Hard/smooth comparisons instead have five location-specific pairs per endpoint, with Holm correction and 99% intervals. The two-factor repeated-measures model excludes the control and does not replace the direct comparisons. Difference directions for those raw direct contrasts are stated explicitly; lower ECE is favorable.

The analysis preserves equivalent, unsupported, and small statistically detectable outcomes without converting one into another. For a constant seed-difference vector, the directional routine returns

$p = 1$ at the tested boundary, $p = 0$ when the constant lies strictly in the stated alternative, and $p = 1$ otherwise. Hence an all-zero vector tested against the strict bounds -0.01 and 0.01 has $p_L = p_U = 0$; its two-sided difference test separately has $p = 1$ and a zero-width interval. These software conventions handle a degenerate observed sample and do not prove population-wide exact equality. Repeated examples, tokens, attack restarts, and timing blocks are not additional model seeds.

A pre-submission statistical audit corrected the multiplicity construction for equivalence: the earlier analysis adjusted the two directional TOST families separately, whereas the current analysis applies Holm to the five location-level intersection–union p-values. The raw paired seed differences, equivalence margins, and directional p-values are unchanged. One decision changes under the corrected procedure: smooth attention-output attacked ECE moves from equivalent to unresolved.

B.7 Supporting evidence and its limits

QKNorm uses five matched classifier seeds, per-head Q/K ℓ_2 normalization, and one learned scale per layer (Henry et al., 2020). The three-seed comparator is *FreeLB-style training with fixed-dropout inner-loop replay*: three inner steps at $0.05r$, step size $\epsilon/3$, and one-third loss weight per inner step, followed by one optimizer update (Zhu et al., 2020). This is a FreeLB-style implementation, not an exact reproduction of every original setting. These baselines use ordinary paired 95% contextual intervals and remain outside the primary multiplicity families. The stronger-attack qualification for FreeLB-style at $0.10r$ remains separate from its primary-budget estimate.

The two TextFooler cap families were selected using validation evidence before official evaluation; the same 250 examples were evaluated for each of their five seeds and the five controls (Jin et al., 2020). Algorithmic success remains separate from human validity. Five consenting reviewers, all serving solely as reviewers, assessed sampled successful attacks independently under blinded conditions. Because the sample is success-conditioned and response coverage varies, the analysis reports reviewer-specific judgments rather than a pooled population-wide human-valid attack-success estimate. Unrecorded changed-word or similarity magnitudes are not imputed.

Historical mechanism measurements come from the training-derived 256-example subset at selected checkpoints. Relative tail compression is $1 - q^{\mathrm{post}}_{.999}/q^{\mathrm{pre}}_{.999}$ for positive pre-cap $q_{.999}$, accompanied by an absolute reduction and an explicit undefined `zero_baseline` state when needed. The source analysis centers values within seed and calculates rank correlations on 50 seed/configuration rows that share only five model seeds. Seed-cluster bootstrap intervals and leave-one-location-out checks are separate exploratory diagnostics. Conventional row-level Spearman p-values remain in the numerical record but are not used for a confirmatory mechanism claim because they do not account for this clustering. The associations remain supporting and noncausal.

Efficiency comes from historical fixed-input microbenchmarks on an NVIDIA A100 graphics processing unit (GPU) with 80 GB of memory: batch 64, length 256, BF16, eager attention, no compilation, 20 warmups, and 50 synchronized timed batches in each of five blocks. Block-level uncertainty is not independent seed replication. Peak-memory scalar ratios do not acquire a formal repeated-block interval when the underlying peak series was not retained. No campaign wall time is substituted for model throughput.

B.8 Evidence chronology and validity-based exclusions

The original design and statistical contracts preceded official-test access; selected checkpoints and cap artifacts have immutable identities. The recovered first-member-request record places the official IMDb test request at 2026-09-03 18:34:22.868883 Coordinated Universal Time (UTC). The prior execution envelope records authorization for one campaign. These records establish an access chronology, not an external preregistration or independent certification of every subsequent operation.

The execution was not a single uninterrupted run of an unchanged runner. Post-access remediation records document failures in first-access request-identity handling and worker tensor-device routing, with the affected attempts retained. Those two records report zero completed scientific outputs at their respective repair boundaries and unchanged frozen scientific settings. Their source changes concern execution, but occurred after test access. Separately, 20 historical attempts lacked suffi-

cient attribution and are excluded from the accepted operation set. The accepted set contains 560 operations, including evidence joins and the invalid language-model component; it is not a count of independent valid model evaluations.

Two later checks change how the records may be interpreted. First, source-row, label, and retained-input matching establish that all 200 strong reviews belong to the primary 1,000. Different selection salts generated different identifier namespaces; the earlier disjoint-subset description is superseded. The recovered comparisons in Table 11 use those exact common source reviews. Second, all 36 official language-model operations used classifier-tokenizer encodings rather than the required Generative Pre-trained Transformer 2 (GPT-2) encodings. Correct token counts and completed execution do not restore benchmark validity. Their NLL, perplexity, attack effects, location tests, and cross-domain replication interpretations are excluded from this paper. The defect was identified after execution; it is neither a prespecified omission nor an intentionally different architecture. Historical training/validation records do not supply official-test confirmation.

Original measurements, subsequent corrections to their interpretation, and privacy-redacted copies have distinct roles in the evidence. Source identifiers retained in the supplementary data refer to the underlying records and need not equal checksums of the released files. The reported classifier analyses use the documented campaigns and statistical families; no corrected official language-model evaluation or cross-domain replication claim is included. Interpretation therefore depends on the stated populations, chronology, and scientific validity of the inputs, rather than on the number of completed operations or passed software checks.

C Complete results and supporting evidence

The main result rests on official-test endpoints and matched model seeds. This appendix makes their variation inspectable and keeps each supporting measurement in its proper role: diagnostic budgets contextualize the primary threat model, contextual baselines compare different training or attention choices, and historical measurements describe activity and cost. The stored values, source identities, and full-precision analysis outputs accompany the project as CSV files. Display rounding does not change a decision. The invalid official language-model computations are excluded from benchmark results throughout.

C.1 The primary result at seed resolution

Clean-utility preservation is visible at the checkpoint level, not only in the aggregate. Table 5 contains all 55 clean-BA values underlying Table 1. Hard MLP is below its matched control in every seed, illustrating why a directional loss and equivalence within a practical margin are compatible. The seed labels preserve the pairing used in inference; treating the 25,000 reviews as independent model replications would answer a different question. Smooth centered logits retain their exact clean-score match to control.

Table 5: Clean balanced accuracy for every primary seed/configuration pair (percent scale). Columns are the five model seeds; rows retain the fixed design order. Each value is linked to the selected checkpoint in the companion CSV. Evaluation uses 25,000 full-clean test reviews. These are individual seed values, not confidence limits.

Configuration	11	23	47	89	131
Uncapped control	92.588	92.360	92.496	92.392	92.016
Hard Embedding	92.652	92.224	92.496	92.404	92.136
Hard Q/K	92.540	92.204	92.320	91.804	91.932
Hard Attention output	92.652	92.352	92.392	92.176	92.052
Hard MLP	91.868	91.976	91.584	91.536	91.600
Hard Centered logits	92.580	92.324	92.460	92.368	92.000
Smooth Embedding	92.432	92.176	92.508	92.248	92.100
Smooth Q/K	92.424	92.424	92.504	92.316	92.124
Smooth Attention output	92.696	92.224	92.520	92.068	92.036
Smooth MLP	92.444	92.188	92.008	91.704	91.816
Smooth Centered logits	92.588	92.360	92.496	92.392	92.016

The apparent advantage of hard attention output also needs this seed-level view. Its primary robust BA ranges from 0.50% to 3.10%, whereas control ranges from 1.30% to 1.90% (Table 6). The wider variation helps explain why its largest positive mean effect does not resolve a practical improvement in Figure 2(b). Both favorable and unfavorable seed outcomes remain in the table; none is selected or discarded on the basis of attack performance.

Table 6: Primary robust balanced accuracy for every primary seed/configuration pair (percent scale). Columns are the five model seeds; rows retain the fixed design order. Each value is linked to the selected checkpoint in the companion CSV. Evaluation uses 1,000 primary attacked reviews at $0.05r$. These are individual seed values, not confidence limits.

Configuration	11	23	47	89	131
Uncapped control	1.600	1.900	1.300	1.500	1.500
Hard Embedding	1.100	1.700	0.600	1.300	1.300
Hard Q/K	0.400	0.900	0.300	0.900	1.600
Hard Attention output	2.300	1.900	0.500	1.900	3.100
Hard MLP	0.400	0.600	0.700	0.800	0.900
Hard Centered logits	1.500	1.900	1.700	1.400	1.200
Smooth Embedding	2.100	1.300	1.100	1.800	1.400
Smooth Q/K	1.600	0.700	1.100	0.900	1.400
Smooth Attention output	1.900	1.000	0.700	2.000	1.600
Smooth MLP	0.500	0.400	0.500	1.000	1.400
Smooth Centered logits	1.600	1.900	1.300	1.500	1.500

Table 7 gives raw attacked ECE for the same 1,000-review primary attack at each checkpoint. The high errors show that low attacked accuracy is not accompanied by well-calibrated confidence. The near-coincident smooth-logit values are retained at their source precision in the CSV, even where table rounding makes them indistinguishable. These are calibration errors on a percent scale, not confidence probabilities. The clean and attacked tables have different populations; a matched attack-induced loss must instead use clean predictions on the attacked reviews.

Table 7: Raw primary attacked ECE-15 for every primary seed/configuration pair (percent scale). Columns are the five model seeds; rows retain the fixed design order. Each value is linked to the selected checkpoint in the companion CSV. Evaluation uses 1,000 primary attacked reviews at $0.05r$. These are individual seed values, not confidence limits.

Configuration	11	23	47	89	131
Uncapped control	97.702	97.639	97.711	96.943	97.769
Hard Embedding	98.438	97.517	98.633	97.455	97.306
Hard Q/K	99.157	98.638	98.955	98.165	97.670
Hard Attention output	97.168	97.518	98.500	96.290	95.404
Hard MLP	98.833	98.804	98.655	98.223	98.105
Hard Centered logits	97.847	97.674	97.176	97.009	98.006
Smooth Embedding	97.340	97.928	97.953	96.806	97.355
Smooth Q/K	97.966	98.598	98.261	97.266	97.115
Smooth Attention output	97.372	98.191	98.084	96.268	96.609
Smooth MLP	99.060	98.928	98.963	98.222	97.538
Smooth Centered logits	97.702	97.639	97.711	96.942	97.769

C.2 Interpreting equivalence and direct comparisons

The E/U annotations in Table 1 and Figure 2 use the corrected location-level TOST decisions. For each location, the equivalence p-value is $\max(p_L, p_U)$; Holm adjustment is then applied across the five locations in that operator/endpoint family. The 99% difference intervals instead provide the stated simultaneous coverage within that family. They need not lie inside ± 1 pp for an equivalence decision to pass. Hard MLP provides a concrete example: its clean difference interval lies below zero, yet the clean effect satisfies the registered equivalence test. Conversely, "unresolved" does not

establish equivalence, benefit, or harm. The complete 30-row records preserve the two one-sided adjusted p-values as well as the separate difference-test results.

None of those cap/control error bars is itself a direct location comparison. The six Greenhouse–Geisser-adjusted location tests detect variation, but none of the 60 direct contrasts satisfies the combined adjusted-significance, simultaneous-interval, and practical-effect rule. The 15 hard-versus-smooth contrasts likewise establish no practical operator advantage under their rule. The accompanying `location_global.csv`, `location_pairwise.csv`, and `hard_vs_smooth.csv` retain every comparison. Their raw $a - b$ or hard-minus-smooth sign convention must not be confused with the favorable ECE orientation in Figure 2(c). Statistical location variation and absence of a demonstrated practical winner are therefore distinct conclusions.

C.3 The calibrated intervention, before capped training

The calibrated parameters document what was imposed, independently of the outcome. Tables 8 and 9 contain all 100 seed/site/layer hard thresholds and their corresponding smooth scales. The variation across native tensors is why a single numerical cap height would not define this study. Figure 1(a) locates those tensors; its operator curves in panel (b) illustrate shape only, whereas panel (c) specifies the actual calibration–freezing–training sequence. The two tables contain the fitted parameters, not values read from that schematic.

Table 8: Fixed hard thresholds τ from the existing training-only calibration artifacts. There are 20 identities per seed; Q/K is pooled within each layer. Values are in activation units and rounded only for display. Layers are zero indexed. Full-precision values, sample counts, matching residuals, and hashes remain in `data/threshold_artifacts.csv`. Scales are not tuned on test performance.

Site / layer	11	23	47	89	131
Embedding (global)	1.0106	0.9765	1.0178	1.0134	0.9963
Q/K 0	7.0000	6.9375	6.9063	6.7500	6.8125
Q/K 1	5.8750	6.0313	6.1250	6.0313	6.0000
Q/K 2	5.9375	5.8125	5.7500	5.9688	5.8750
Q/K 3	5.9375	6.0000	6.0000	6.0938	6.0000
Q/K 4	5.9688	6.1250	6.3125	6.1563	6.4375
Q/K 5	5.8438	5.6250	5.7813	6.0625	5.8750
Attention output 0	0.6367	0.6406	0.6289	0.6172	0.6172
Attention output 1	1.2422	1.2891	1.2422	1.1797	1.2734
Attention output 2	1.2578	1.2188	1.2578	1.1484	1.2500
Attention output 3	1.1641	1.1797	1.1563	1.1250	1.1484
Attention output 4	1.3906	1.2656	1.1953	1.1797	1.1719
Attention output 5	2.3281	2.2656	2.1094	1.9297	2.1875
MLP 0	1.2500	1.1953	1.2031	1.2031	1.2266
MLP 1	1.4063	1.3906	1.3281	1.2109	1.3672
MLP 2	1.2656	1.3594	1.3281	1.2891	1.3359
MLP 3	1.3594	1.2891	1.3047	1.2578	1.2656
MLP 4	1.2969	1.3047	1.4297	1.3281	1.4297
MLP 5	1.6563	1.7734	1.6563	1.5547	1.6719
Centered logits (global)	3.0938	3.1875	2.9844	2.6875	2.9219

Every accepted matched scale exceeds its hard threshold, consistent with the consequence of Theorem 1. This means that equal numerical parameters would not have compared equal calibration distortion. It does not mean that either scale is optimal, or that gradients and subsequent engagement were matched. The tables describe the selected control's training-only calibration artifacts; neither official-test outcomes nor later activity measurements select their values. Full-precision entries in `threshold_artifacts.csv`, rather than the rounded tables, are required to check the recorded matching residuals.

Table 9: Distortion-matched smooth scales $s^\star$ from the existing training-only calibration artifacts. There are 20 identities per seed; Q/K is pooled within each layer. Values are in activation units and rounded only for display. Layers are zero indexed. Full-precision values, sample counts, matching residuals, and hashes remain in `data/threshold_artifacts.csv`. Scales are not tuned on test performance.

Site / layer	11	23	47	89	131
Embedding (global)	2.5506	2.6838	2.5946	2.6509	2.6407
Q/K 0	17.6673	17.3789	17.5950	16.7796	17.4419
Q/K 1	19.5923	20.0507	20.6277	20.7197	20.6769
Q/K 2	23.6629	23.0146	21.9684	24.8010	22.7471
Q/K 3	26.5558	26.6300	26.8516	26.4332	25.6311
Q/K 4	25.7187	24.5292	26.9252	26.1150	26.1935
Q/K 5	22.4355	21.4262	21.1863	20.9926	21.9769
Attention output 0	2.2175	2.1313	2.0814	2.0204	2.0126
Attention output 1	4.1247	4.2670	4.0399	3.7811	4.0912
Attention output 2	4.3147	4.3568	4.5943	4.0055	4.3689
Attention output 3	4.5295	4.3364	4.5125	4.5282	4.3958
Attention output 4	6.4723	5.4207	4.6970	4.7621	4.7205
Attention output 5	6.3752	6.1268	5.2381	6.6678	6.3231
MLP 0	2.0498	2.0396	2.0431	2.0709	2.0651
MLP 1	2.4461	2.4779	2.3967	2.1010	2.4165
MLP 2	2.2688	2.4789	2.3140	2.2980	2.3561
MLP 3	2.6733	2.5168	2.6537	2.4288	2.5116
MLP 4	2.3963	2.4239	2.6752	2.5463	2.8173
MLP 5	3.3111	3.6057	3.2167	2.8565	3.1028
Centered logits (global)	152.1763	216.4342	120.3623	143.4287	176.1826

C.4 Post-hoc sensitivity at the non-saturating budget

Table 10 reports the complete $0.025r$ paired analysis specified in Appendix B.6. The control mean is 28.78% robust BA. No capped configuration has a Holm-corrected positive difference. Hard Q/K is lower by 5.56 pp and its 99% interval excludes zero; hard and smooth centered logits satisfy the corrected equivalence test. The other seven contrasts remain unresolved, including the $+1.62$ pp hard attention-output mean. These post-hoc results address the primary floor with the existing five-seed experiment without redefining the registered endpoint or selecting a favorable budget.

Table 10: Post-hoc $0.025r$ robust-BA sensitivity analysis on the original 1,000-review attack population. The control mean is 28.78%. Effects and paired SDs are percentage points over five matched seeds. Intervals are individual 99% paired-t intervals, providing at least 95% Bonferroni coverage across the five locations within each operator family under the paired-t assumptions. p_{diff} is the within-operator Holm-adjusted two-sided difference test; p_{eq} is Holm-adjusted after forming the location-level intersection–union equivalence p-value. The equivalence margin is ± 1 pp. This post-hoc analysis does not replace the prespecified $0.05r$ endpoint or assess the full multi-endpoint benefit rule.

Site	Cap BA	Δ	Paired SD	99% interval	p_{diff}	p_{eq}	Decision
Hard cap							
Embedding	27.30	-1.48	2.48	[-6.58, +3.62]	0.579	1.000	Unresolved
Q/K	23.22	-5.56	2.32	[-10.34, -0.78]	0.029	1.000	Lower
Attention output	30.40	+1.62	2.32	[-3.15, +6.39]	0.579	1.000	Unresolved
MLP	24.42	-4.36	2.58	[-9.67, +0.95]	0.078	1.000	Unresolved
Centered logits	28.74	-0.04	0.15	[-0.35, +0.27]	0.587	$< .001$	Equivalent
Smooth cap							
Embedding	28.70	-0.08	1.70	[-3.59, +3.43]	1.000	0.587	Unresolved
Q/K	28.98	+0.20	1.52	[-2.94, +3.34]	1.000	0.587	Unresolved
Attention output	29.20	+0.42	2.91	[-5.58, +6.42]	1.000	0.679	Unresolved
MLP	24.42	-4.36	3.90	[-12.39, +3.67]	0.334	0.937	Unresolved
Centered logits	28.78	+0.00	0.00	[+0.00, +0.00]	1.000	$< .001$	Equivalent

C.5 MATCHED OPTIMIZATION AND BACKWARD-PASS DIAGNOSTICS

Table 11 exposes the starting and ending accuracies behind Figure 3(c). Both optimization budgets use the same 200 source-matched reviews at each checkpoint. Every primary 20×3 starting mean is at most 1%, so an absolute decrease strictly greater than 2 pp is impossible by the metric's lower bound. The registered warning threshold is therefore non-diagnostic for those rows: small observed decreases neither prove under-optimization nor certify convergence. QKNorm is already at zero in both routes, so its zero decrease cannot be taken as evidence of robust predictions.

FreeLB-style illustrates why the actual accuracy levels and budget matter. On this 200-review subset, its three-seed mean changes from 57.67% to 57.33% at $0.05r$, but from 11.33% to 8.67% at $0.10r$ (Table 11). All three stress-budget seed decreases exceed 2 pp; their stored mean is 2.667 pp. This warning qualifies the stress-budget attack estimate, not the separate primary-budget effect against control. The 26 configuration summaries and all 126 paired seed/budget records are retained, rather than treating the summaries as independent experimental units.

Table 11: Source-matched stronger-attack diagnostics. Means are over five seeds for primary configurations and QKNorm, and three for FreeLB-style. Both optimizers use the same 200 nested test reviews per checkpoint. BA is percent; decrease is percentage points. Every primary starting mean is at most 1%, so a drop strictly greater than 2 pp is impossible and the warning threshold is non-diagnostic for those rows. FreeLB-style exceeds 2 pp at $0.10r$, not at $0.05r$.

	$0.05r$			$0.10r$		
Configuration	20×3	50×10	Decrease	20×3	50×10	Decrease
Uncapped control	1.00	0.80	0.200	0.00	0.00	0.000
Hard Embedding	0.80	0.60	0.200	0.10	0.00	0.100
Hard Q/K	0.40	0.40	0.000	0.00	0.00	0.000
Hard Attention output	0.90	0.90	0.000	0.10	0.00	0.100
Hard MLP	0.10	0.10	0.000	0.00	0.00	0.000
Hard Centered logits	1.00	0.80	0.200	0.00	0.00	0.000
Smooth Embedding	0.80	0.80	0.000	0.00	0.00	0.000
Smooth Q/K	1.00	1.00	0.000	0.00	0.00	0.000
Smooth Attention output	1.00	0.80	0.200	0.10	0.10	0.000
Smooth MLP	0.30	0.20	0.100	0.00	0.00	0.000
Smooth Centered logits	1.00	0.80	0.200	0.00	0.00	0.000
QKNorm	0.00	0.00	0.000	0.00	0.00	0.000
FreeLB-style	57.67	57.33	0.333	11.33	8.67	2.667

Table 12 answers a separate question: with the exact hard-cap forward map fixed, what changes when the attack uses native derivatives rather than identity BPDA? Positive entries mean BPDA obtains lower accuracy. Most entries are zero; the positive differences are concentrated at centered logits, with one negative Q/K entry also retained. This distinguishes backward-route sensitivity from the step/restart comparison above. Proposition 3 concerns the exact boundary of a fixed model, not whether two finite-step gradient searches return the same candidate. No numerical native/BPDA materiality threshold was prespecified, and these finite searches cannot establish a universal absence of gradient masking.

C.6 CONTEXTUAL BASELINES AND PROBABILITY CALIBRATION

The contextual baselines show that scale control and adversarial training are not interchangeable interventions. QKNorm's five-seed robust-BA effect against its own matched control is -1.440 pp, with the ordinary paired 95% interval $[-1.993, -0.887]$ pp (Table 13). Its clean BA and raw attacked ECE, NLL, and Brier effects are also unfavorable. These observations characterize this QKNorm implementation and classifier setting; they do not establish that normalization is generally harmful or that a particular cap is formally superior to QKNorm.

FreeLB-style has a different profile in Table 13. Alongside its large primary robustness gain, attacked ECE, NLL, and Brier improve relative to its three matched controls. Its clean-BA effect is -0.359 pp, with the ordinary 95% interval $[-0.924, +0.206]$ pp. These contextual estimates remain

Table 12: Native-hard minus BPDA balanced accuracy on the common 200-example stronger-attack subset, in percentage points. All 25 hard checkpoints and both budgets are included. Positive values mean that BPDA obtained lower accuracy. This is not the 20×3-versus-50×10 comparison, and no separate numerical native/BPDA materiality threshold was prespecified.

Hard-cap site	Seed	$0.05r$	$0.10r$
Embedding	11	0.00	0.00
Embedding	23	0.00	0.00
Embedding	47	0.00	0.00
Embedding	89	0.00	0.00
Embedding	131	0.00	0.00
Q/K	11	0.00	0.00
Q/K	23	0.00	0.00
Q/K	47	0.00	0.00
Q/K	89	-0.50	0.00
Q/K	131	0.00	0.00
Attention output	11	0.00	0.00
Attention output	23	0.00	0.00
Attention output	47	0.00	0.00
Attention output	89	0.00	0.00
Attention output	131	0.00	0.00
MLP	11	0.00	0.00
MLP	23	0.00	0.00
MLP	47	0.00	0.00
MLP	89	0.00	0.00
MLP	131	0.00	0.00
Centered logits	11	+0.50	+1.00
Centered logits	23	+0.50	+1.00
Centered logits	47	+1.50	+2.00
Centered logits	89	0.00	0.00
Centered logits	131	+0.50	+0.50

separate from the primary capping multiplicity families. They show that the evaluator distinguishes at least one substantially more robust system, without attributing the improvement to an individual training ingredient or equalizing training cost. The stronger-attack warning at $0.10r$ remains as stated in Table 11.

Table 13: Contextual baseline effects relative to same-seed controls. Every effect is oriented so that positive is favorable. Intervals are the stored ordinary paired 95% intervals, not the multiplicity-adjusted primary intervals. QKNorm uses five seeds; FreeLB-style uses three. Robustness and raw attacked calibration use the primary $0.05r$ setting. These are not compute-matched cap-versus-baseline superiority tests.

Method	Endpoint / units	Seeds	Effect	Ordinary 95% CI
QKNorm	Clean BA (pp)	5	-0.788	[-1.131, -0.445]
QKNorm	Robust BA (pp)	5	-1.440	[-1.993, -0.887]
QKNorm	Favorable ECE (pp)	5	-2.075	[-2.620, -1.530]
QKNorm	Favorable NLL (nats)	5	-0.730	[-0.977, -0.482]
QKNorm	Favorable Brier	5	-0.047	[-0.059, -0.035]
FreeLB-style	Clean BA (pp)	3	-0.359	[-0.924, +0.206]
FreeLB-style	Robust BA (pp)	3	+59.333	[+58.329, +60.337]
FreeLB-style	Favorable ECE (pp)	3	+74.977	[+73.208, +76.745]
FreeLB-style	Favorable NLL (nats)	3	+4.841	[+4.276, +5.406]
FreeLB-style	Favorable Brier	3	+1.394	[+1.377, +1.411]

Figure 4(a) compares each checkpoint's raw full-clean ECE with the ECE obtained using its stored validation-fitted temperature. Points below the identity line have lower error after that fixed scaling. All 55 checkpoint pairs are present, including overlaps; the line is a reference, not a fitted regression.

This asks how the recorded probabilities respond to a stored calibration transformation, not whether test-set refitting can improve them.

Panel (b) makes the distinction from robustness visible: the attacked ECE remains high even where scaling reduces it. Both coordinates refer to the same checkpoint and the same 1,000 attacked examples. The panel has its own axis range and must not be visually compared with the clean panel as though their scales were equal. Supporting natural-log NLL and the summed two-class Brier score are preserved in `calibration_sensitivity_per_seed.csv`. Each scaled record is matched to its validation-fitted temperature by checkpoint and exact value when its embedded identity field is absent. Temperature is not fitted on test data; raw attacked ECE remains the primary endpoint in Figure 2(c).

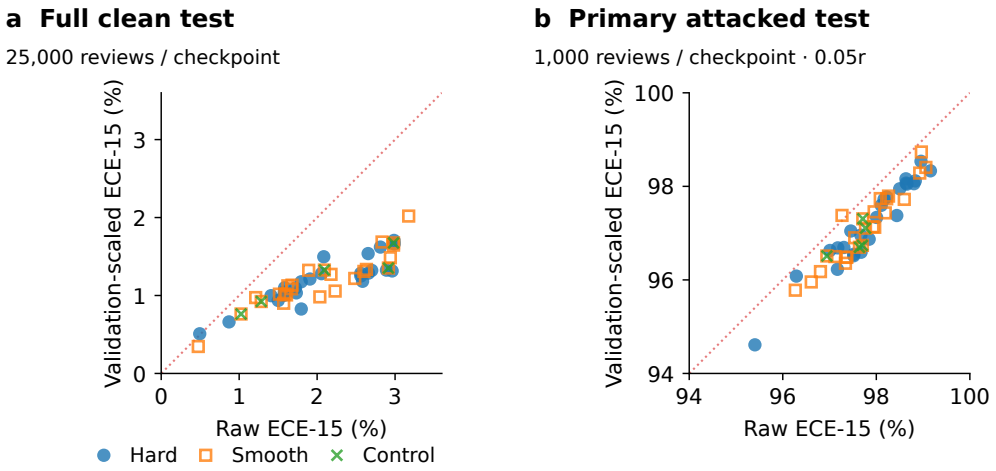

Figure 4: **Temperature sensitivity is not a replacement endpoint.** Each panel contains all 55 primary seed/checkpoint pairs, pairing raw ECE with the same checkpoint's stored validation-temperature-scaled ECE. (a) Full clean test: 25,000 reviews per checkpoint. (b) Primary attacked test: 1,000 reviews at $0.05r$. Both axes use a percent scale; panel ranges differ. Points below the identity line have lower ECE after scaling. No test-set temperature is fitted, no regression is added, and overlapping points are not removed. Raw attacked calibration remains primary. Temperature provenance retains the documented checkpoint/value join where an embedded identity field was null.

C.7 Historical mechanisms and efficiency

Historical measurements provide an exploratory view of the representation changes, not a second official-test experiment. Table 14 retains all six rank correlations and their seed-cluster bootstrap intervals. Greater $q_{.999}$ tail compression is directionally associated with less favorable clean, robustness, and calibration effects, whereas the exceedance-rate associations are smaller. The mechanisms were measured on the training-derived subset and then joined to official endpoint effects. The 50 seed/configuration rows contain only five independent model seeds; with so few clusters, the intervals are descriptive stability summaries rather than confirmatory mechanism inference. Conventional row-level Spearman p-values and their Holm corrections remain in the numerical records but are not used here to support significance. These associations neither identify a causal effect of tail suppression nor carry the primary robustness conclusion.

Figure 5 shows that the local operators also have location-dependent recorded costs. Smooth MLP and smooth Q/K have lower throughput than their matched controls; other sites include smaller losses or gains. All 50 capped checkpoints and eight contextual checkpoints remain visible, so the plot is not a selection of the slowest methods. Its intervals describe the five timing blocks within each checkpoint, not uncertainty over independent model seeds or future hardware. Reference lines at $\pm 5\%$ mark the historical practical scale; percentage throughput changes are not percentage-point accuracy effects.

Table 14: Exploratory historical-mechanism associations with official classifier effects. The 50 seed/configuration rows comprise only five independent model seeds. The seed-cluster bootstrap intervals summarize directional stability but, with five clusters, do not support confirmatory mechanism inference. Conventional row-level Spearman p-values are retained in the numerical record but omitted here because they are not cluster-aware. All six recorded associations are shown; no causal or new official-test mechanism claim is made.

Historical variable	Endpoint	ρ	Seed-cluster bootstrap interval
Exceedance rate	Clean BA	-0.249	[-0.390, -0.165]
Exceedance rate	Robust BA	-0.068	[-0.281, +0.115]
Exceedance rate	Favorable ECE	-0.123	[-0.291, +0.072]
Tail compression	Clean BA	-0.415	[-0.555, -0.260]
Tail compression	Robust BA	-0.488	[-0.580, -0.367]
Tail compression	Favorable ECE	-0.502	[-0.675, -0.280]

These are fixed-input historical microbenchmarks, not official-test campaign wall time (Figure 5). The accompanying CSV preserves the recorded hardware/runtime comparison and missing-field qualifications; unrecorded runtime fields are not reconstructed. Scalar peak-memory ratios are retained, but no repeated-block peak series exists from which to report a corresponding formal interval. Throughput costs can therefore be discussed at the scope of the recorded microbenchmark without suggesting a universal runtime or memory guarantee.

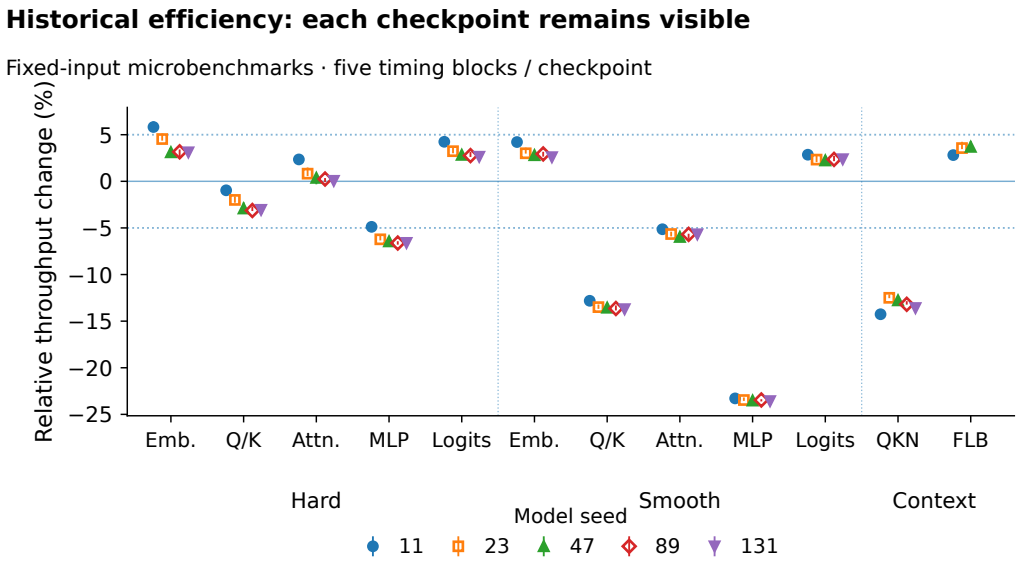

Figure 5: **Historical throughput, with every checkpoint retained.** Relative change against the matched control, in percent (not percentage points), for 50 capped and eight contextual checkpoints. Markers distinguish model seeds; each interval is the retained bootstrap interval over five timing blocks within that checkpoint. These are not confidence intervals over independent model seeds. Reference lines mark zero and $\pm 5\%$. Measurements use historical fixed-input, batch-64, length-256, BF16 eager microbenchmarks, not official-test campaign wall time. QKN and FLB denote QKNorm and FreeLB-style. Retained runtime-compatibility qualifications and scalar peak-memory ratios remain in the CSV; no unsupported memory interval is drawn.

C.8 Discrete attacks and reviewer-specific validity

The discrete-attack evidence asks a different question from the continuous embedding-space evaluation. Table 15 contains the complete TextFooler outcome accounting for control and the validation-selected hard attention-output and smooth-embedding families: 3,123 successes, 152 unsuccessful

attacks, 204 query-exhausted cases, and 271 clean-incorrect cases. Success among clean-correct evaluations is approximately 89–90% in each family. This similarity is descriptive, not a new equivalence result. The same 250 source reviews are reused across 15 checkpoints; 3,750 model–example evaluations do not supply 3,750 independent reviews or model seeds. Neither algorithmic success nor a stored quality flag establishes human validity.

Table 15: TextFooler algorithmic outcomes on the fixed 250-review subset, evaluated at five selected checkpoints per family. Each row comprises 1,250 model–example evaluations, not 1,250 independent reviews or seeds. The success percentage uses clean-correct examples only. Automatic success and a stored quality flag are not human validity.

Family	Success / clean-correct	%	Failed	Exhausted	Clean-wrong
Control	1045/1156	90.40	51	60	94
Hard Attention output	1043/1165	89.53	48	74	85
Smooth Embedding	1035/1158	89.38	53	70	92

Table 16 separates sentiment preservation from grammatical acceptability for all five reviewers. It displays the actual yes/nonmissing counts for each family and question, including incomplete submissions rather than discarding or filling them. Sentiment judgments are generally favorable, whereas grammar judgments are lower and strongly reviewer-dependent: even the two complete submissions give markedly different acceptance counts. Uncertain responses remain in the stated nonmissing denominator.

Chance-adjusted agreement is summarized by nominal Krippendorff's alpha (Hayes & Krippendorff, 2007). Across the 150 items with at least two ratings, $\alpha = 0.153$ for sentiment preservation and $\alpha = 0.371$ for grammatical acceptability. The four uncertain sentiment responses are treated as a separate nominal category, missing responses are omitted, and these post-collection coefficients are descriptive diagnostics rather than acceptance gates.

These judgments concern 150 sampled *successful-attack* pairs, comprising 111 distinct source reviews, not a random sample of all attack attempts (Table 16). Each family's selected successes need not be the same attacks, so their descriptive ordering does not establish a capping-specific grammar advantage. All five reviewers consented, served solely as reviewers, and submitted ratings independently under blinded conditions. Two reviewers rated all 150 pairs, two rated 137, and one rated 41; missing responses remain missing. Exposure time was not recorded, and R1–R5 are reviewer aliases. No consensus, adjudication, imputation, or pooled human-valid attack-success rate is introduced. These success-conditioned judgments qualify the interpretation of automatic attack success without estimating population-wide human-valid attack success.

Table 16: Reviewer-specific yes/nonmissing responses, conditioned on 150 sampled successful-attack pairs (50 per model family; 111 distinct source reviews). H denotes selected hard attention output; S denotes selected smooth embedding. Uncertain responses remain in the nonmissing denominator; missing responses are not imputed. All five consenting reviewers served solely as reviewers and submitted ratings independently under blinded conditions. Exposure time was not recorded. The counts describe the sampled successful attacks, not pooled population-wide human-valid attack-success rates.

	Sentiment preserved			Grammar acceptable		
Reviewer	Control	H	S	Control	H	S
R1	49/50	49/50	50/50	9/50	16/50	12/50
R2	49/50	49/50	50/50	2/50	8/50	4/50
R3	42/45	42/46	45/46	9/45	11/46	13/46
R4	45/45	45/46	45/46	0/45	0/46	0/46
R5	14/14	17/17	10/10	0/14	4/17	3/10

D ADDITIONAL LITERATURE CONTEXT AND COMPARISON BOUNDARIES

Modified rectifiers, learned caps, and representation attacks. The modified-ReLU study of Sooksatra et al. (2024a) and the later learned maxima of Sooksatra & Rivas (2024) concern nonnegative rectifiers, not the fixed symmetric operators compared here. The latent-text study of Sooksatra et al. (2024b) perturbs representations and maps candidates back to text. That distinction matters because success in a continuous representation space does not, by itself, establish a valid discrete textual attack. Sooksatra et al. (2023) instead study adversarial generators against reconstruction models. Together, these studies distinguish three attack targets: internal representations, discrete text, and reconstructed outputs. Our primary endpoint concerns classifier decisions under the specified representation perturbations; the secondary textual evaluation addresses a separate validity question.

Outliers, quantization, normalization, and alignment. Clipped softmax and gated attention were proposed to reduce outliers and improve quantization (Bondarenko et al., 2023); SmoothQuant uses a mathematically equivalent rescaling to move quantization difficulty from activations to weights (Xiao et al., 2023). These procedures differ from coordinate clipping. Massive-activation studies further identify functional roles for large internal values (Sun et al., 2024), but that literature does not establish the cause of our historical tail-compression associations. Input-conditioned Dynamic ReLU (Chen et al., 2020) is also distinct from Dynamic-Max-Value ReLU. More recently, coupled weight and activation constraints address safety drift during language-model fine-tuning (Peng et al., 2026). Our binary sentiment-classification experiment provides no refusal, persona, alignment, or general language-model safety result.

Attention regularization and certification. ProTransformer changes attention through a plug-and-play construction (Hou et al., 2024), whereas Attention-Regularized Transformers (ART) regularize attention sensitivity during training (Bouri et al., 2026). Learned clipping also appears within network-level local-Lipschitz certification (Huang et al., 2021). Such systems impose conditions or alter optimization beyond the fixed coordinate caps studied here. Their results cannot be inferred from our local nonexpansiveness proposition, and our empirical comparisons do not refute those distinct procedures.

Training objectives and uncertainty. The robust fine-tuning method SMART regularizes local behavior through smoothness-inducing adversarial regularization and proximal optimization (Jiang et al., 2020); InfoBERT adds information-bottleneck and robust-feature regularizers to language-model fine-tuning (Wang et al., 2021). Confidence-calibrated adversarial training also changes the training objective and supports rejection of low-confidence inputs (Stutz et al., 2020), whereas RanMASK certifies robustness through randomized input masking (Zeng et al., 2023). These methods alter training or inference beyond the fixed coordinate caps studied here and are not omitted instances of our intervention. Temperature scaling follows calibration methodology (Guo et al., 2017), not a proof of adversarial correctness. BERT-ATTACK supplies a related masked-language-model substitution approach (Li et al., 2020); our discrete evaluation instead uses the frozen TextFooler recipe. For statistical interpretation, the TOST construction (Schuirmann, 1987; Lakens, 2017) and sequential multiplicity adjustment (Holm, 1979) provide methodological references. Our analysis retains the prespecified margins, families, and seed-level inferential unit; an ordinary unadjusted equivalence interval is not the corrected decision rule used here.

E REPRODUCIBILITY AND EVIDENCE PROVENANCE

The supplementary data provide the numerical results, per-seed measurements, attack diagnostics, calibrated thresholds, and evaluation-population definitions underlying the reported findings. They complement the methods, proofs, and experimental details in the paper, allowing readers to inspect the comparisons beyond the precision of the printed tables and figures.

E.1 NUMERICAL RESULTS AND EXPERIMENTAL SETTINGS

The results in Table 1 and Figure 2 are provided in `primary_table.csv`, `primary_effects.csv`, `primary_per_seed.csv`, and `primary_paired_effects.csv`, within

the supplementary `data/` directory. These files preserve the stored numerical precision, matched-seed values, effect estimates, uncertainty, and recorded decisions. The equivalence, noninferiority, and direct-comparison files retain their respective inferential families and effect directions. `units_and_intervals.csv` distinguishes fractions, percentages, percentage-point differences, sample standard deviations, paired-difference intervals and their family-wise coverage, ordinary contextual intervals, and within-checkpoint timing-block intervals. These distinctions are essential when comparing results across endpoints and evaluation procedures.

Figure 3 draws on two distinct evaluation populations. `budget_summary.csv` and `budget_per_seed.csv` contain the primary classifier results at all three measured attack budgets, and `lower_budget_sensitivity.csv` contains the post-hoc $0.025r$ paired analysis. `matched_strong_summary.csv` and `matched_strong_per_seed.csv` contain the source-matched comparisons between 20-step, three-restart attacks and 50-step, ten-restart attacks on the same 200-review subset. `native_bpda_per_seed.csv` separately records the comparison between native hard-cap gradients and BPDA. The operator curves in Figure 1 illustrate the functions rather than fitted experimental parameters; the calibrated thresholds and smooth scales are provided in `threshold_artifacts.csv` and summarized in Tables 8–9.

The calibration, mechanism, efficiency, and reviewer summaries preserve the populations and measurement stages described in Appendix C. In particular, historical mechanism and timing measurements remain distinct from official-test classifier outcomes, and success-conditioned human ratings remain distinct from attack success over the full evaluation population. `populations.json` records the corresponding population definitions, sample counts, and subset relationships. Source and checkpoint identifiers associate numerical entries with retained records. They preserve provenance, while the study design and validity checks determine the scientific interpretation of those entries.

E.2 Scope of the supplementary data

The release is limited to numerical data and scientific metadata. It does not include figure files, LaTeX table sources, display-generation programs, or operational logs. Model weights, raw review text, token arrays, individual rating forms, and executable research environments are also excluded. Selected checkpoint identities are retained in the relevant numerical files, while Appendix B specifies the calibration, training, and evaluation settings.

The supplementary data support inspection of the reported comparisons, their uncertainty, and their evaluation conditions. They are not a complete software release for rerunning the training and adversarial-evaluation campaign. The mathematical results are self-contained: Theorem 1 has a complete main-text proof and an expanded treatment in Appendix A.

E.3 Human-rating provenance and interpretation

Five reviewers consented to participate, served solely as reviewers, and submitted their ratings independently under blinded conditions. Exposure time was not recorded. The 615 reviewer–item records cover 150 sampled successful-attack pairs and 111 distinct source reviews. Two reviewers rated all 150 pairs, two rated 137, and one rated 41. Missing responses remain missing, and the four uncertain sentiment responses remain uncertain. R1–R5 are reviewer aliases. Nominal Krippendorff's alpha is reported in Appendix C.8, with uncertain sentiment as a separate category and missing responses omitted.

Table 16 reports each reviewer's actual denominators, without consensus or imputation. Because the model-family samples are conditional on successful attacks, their ratings do not estimate human-valid attack success across all attempted attacks. The ratings provide independent, blinded judgments of sentiment preservation and grammatical acceptability for the sampled pairs. They are supporting evidence and do not alter the primary classifier endpoint or its statistical interpretation.